NUWC-NPT Technical Report 11,274
20 March 2001

Theory of Continuous-State Hidden Markov Models and Hidden Gauss-Markov Models

Phillip L. Ainsleigh
Submarine Sonar Department

20030711 049

Naval Undersea Warfare Center Division
Newport, Rhode Island

Approved for public release; distribution is unlimited.

PREFACE

This report was prepared under Project No. A101401, "Automatic Signal Classification Project," principal investigator Stephen G. Greineder (Code 2121). The sponsoring activity is the Office of Naval Research, program manager John Tague (ONR 321).

The technical reviewer for this report was Marcus L. Graham (Code 2211).

The author expresses his gratitude to Tod Luginbuhl, Steve Greineder, Paul Baggenstoss, and Roy Streit of NUWC Division Newport for the many stimulating technical discussions that influenced this work. The author also extends appreciation to his doctoral advisory committee at Texas A&M University (Nasser Kehtarnavaz, Edward Dougherty, Don Halverson, and Emanuel Parzen), all of whom provided invaluable feedback during the preparation of the original thesis from which this report is derived.

Reviewed and Approved: 20 March 2001

R J Martin

Ronald J. Martin
Head, Submarine Sonar Department

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	20 March 2001	

4. TITLE AND SUBTITLE

Theory of Continuous-State Hidden Markov Models and Hidden Gauss-Markov Models

5. FUNDING NUMBERS

6. AUTHOR(S)

Phillip L. Ainsleigh

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Undersea Warfare Center Division
1176 Howell Street
Newport, RI 02841-1708

8. PERFORMING ORGANIZATION REPORT NUMBER

TR 11,274

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

A general theory of continuous-state hidden Markov models is developed, with continuous-state analogs of the Baum, Viterbi, and Baum-Welch algorithms formulated for this class of models. The algorithms are specialized to models with linear Gaussian densities, thereby unifying the theory of hidden Markov models and Kalman filters. The Baum and Viterbi algorithms for Gaussian models are shown to be implemented by two different formulations of the fixed-interval Kalman smoother. Moreover, the measurement likelihoods obtained from the forward pass of the Baum algorithm and from the Kalman-filter innovation sequence are found to be equivalent. A direct link between the Baum-Welch algorithm and an existing expectation-maximization algorithm for linear Gaussian models is demonstrated. The general continuous-state and Gaussian models are extended to incorporate mixture densities for the prior probability of the initial state. For the Gaussian models, a new expression for the cross covariance between time-adjacent states is derived from the off-diagonal block of the conditional joint covariance matrix and a parameter invariance structure is observed when the system matrices are time invariant.

14. SUBJECT TERMS

Baum Algorithm	Kalman Filters	Gauss-Markov Model
Baum-Welch Algorithm	Continuous-State Model	Hidden-Markov Model
Viterbi Algorithm	Gaussian Model	EM Algorithm

15. NUMBER OF PAGES 76

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

TABLE OF CONTENTS

Section		Page
	LIST OF TABLES	ii
	FOREWORD	iii
1.	INTRODUCTION	1
1.1	Background	1
1.2	Related Literature	3
1.3	Original Contributions	5
2.	CONTINUOUS-STATE HIDDEN MARKOV MODELS	9
2.1	Baum Probability Densities	11
2.2	Likelihood Evaluation	15
2.3	Viterbi Algorithm	15
2.4	Parameter Estimation	16
2.5	Mixed-Mode Models	20
3.	HIDDEN GAUSS-MARKOV MODELS	25
3.1	Gaussian Refactorization Lemma	26
3.2	Baum Algorithm	27
3.3	Viterbi Algorithm	36
3.4	Parameter Estimation	37
3.5	Time-Invariant Models	44
3.6	Mixed-Mode Models	45
4.	SUMMARY	47
	APPENDIX A - PROOF OF THE REFACTORIZATION LEMMA	49
	APPENDIX B - HGMM JOINT STATE DENSITY	55
	APPENDIX C - HGMM PARAMETER INVARIANCE PROPERTY	59
	REFERENCES	63

LIST OF TABLES

Table		Page
1	Notation for Hidden Markov Models	10
2	Baum Probability Functions	12
3	Baum Algorithm for DS-HMMs	13
4	Baum Algorithm for CS-HMMs	16
5	EM Auxiliary-Function Components for CS-HMMs	20
6	Baum Algorithm for Mixed-Mode CS-HMMs	23
7	Baum Algorithm for HGMMs	33
8	EM Parameter Estimators for HGMMs	43

FOREWORD

The material presented in this report constitutes a portion of the author's doctoral thesis from Texas A&M University, entitled "Segmented Chirp Features and Hidden Gauss-Markov Models for Transient Signal Classification.". In addition to the material presented here, the thesis discusses the feature set used to characterize segments of wandering-tone transients, and it contains a lengthy chapter on simulation and experimental results for classifying these signal types. Since the theoretical developments relating to continuous-state and Gaussian hidden Markov models should be of interest to a much wider audience, it seems desirable to have a self-contained discussion of these topics, without the "baggage" of that additional material. This report provides such a discussion.

Selected portions of this material have also been submitted to the *IEEE Transactions on Signal Processing* in a proposed article entitled "Hidden Gauss-Markov Models for Signal Classification," by P. L. Ainsleigh, N. Kehtarnavaz, and R. L. Streit. The space constraints of such a journal preclude the full development given here, however.

THEORY OF CONTINUOUS-STATE HIDDEN MARKOV MODELS AND HIDDEN GAUSS-MARKOV MODELS

1. INTRODUCTION

Two of the fundamental problems that have driven the development of signal processing algorithms during the past several decades are *tracking*, whose goal is to characterize the time evolution of a "source" through space, frequency, or some other physical variable, and *classification*, whose aim is to make a decision about the nature of a source from its observable characteristics. Tracking can play a significant role in classification since much of the information about the nature of a source can be inferred from the way that it "moves" through the domain of a physical variable. In spite of this, the two fields have historically evolved independently, giving rise to two separate families of tools. From the tracking side comes the Kalman filter and a number of related smoothing algorithms. From the classification side comes the hidden Markov model (HMM) and its associated algorithms. This report unifies these two bodies of theory.

The idea of an equivalence between Kalman filters and HMMs will come as no surprise to many readers, for the analogous nature of these two areas has been noted in the literature for many years. The specifics of the equivalence may be more surprising, however. The Kalman filter is not just analogous to an HMM. The Kalman-filter model *is* an HMM with linear Gaussian model densities. The Baum algorithm for this HMM *is* a Kalman smoother, as is the Viterbi algorithm. The likelihood defined by the HMM criteria is analytically the same as the likelihood defined for a Kalman filter. Each iteration of the expectation-maximization (EM) parameter-estimation algorithm for Kalman-filter models maximizes an auxiliary function whose structure is identical to the function whose maximization gives the Baum-Welch re-estimation formula for HMMs. The equivalences are demonstrated here in great detail, thus removing any ambiguity about the relationship between these two tools.

1.1 BACKGROUND

The material presented in this report resulted from an investigation of methods for classifying certain types of transient signals. In particular, this work is motivated by a persistent difficulty that occurs in the design of statistical signal classifiers. The source of this difficulty is the high dimensionality of feature sets required to adequately describe

signals from each class, which often results in difficult or impossible probability-density estimation problems [1]. An approach for dealing with this issue is to segment signals in time, develop a low-dimensional feature set that provides good signal approximations on the shorter time segments, and use a parameterized stochastic model to evaluate the time evolution of the feature values relative to each class. The objective of such an approach is to replace a high-dimensional time-invariant probability distribution with a low-dimensional time-varying distribution, thereby exchanging the difficult problem of estimating densities in high-dimensional feature spaces for an easier model parameter estimation problem.

In addition to alleviating the dimensionality problem, this type of stochastic feature modeling helps to accommodate the within-class variability that is commonly encountered in real-world signal classes. It is able to do this because the stochastic model allows for "controlled uncertainty" when characterizing the feature evolution, effectively forming an entire family of template-like models that account for the different variations.

Stochastic feature modeling is considered here in the context of maximum-likelihood classification, which is performed in a conventional or a class-specific framework [2]. In the conventional mode, classes are distinguished whose signal segments are all parsimoniously represented by the same set of features. The features are tracked using a parameterized model $\mathcal{M}(\Theta)$. Sequences of feature values from different classes are represented using different values for the parameter set Θ, which are estimated by maximizing the likelihood of a set of labeled sequences (i.e., a training set) from each class. After Θ is estimated for each class, an unlabeled feature sequence $\mathbf{Z}$ is classified by assigning it to the class for which the class-conditional likelihood $\mathcal{L}\{\mathbf{Z}|\Theta\}$ is maximum. In the class-specific framework, distinct feature sets are used to represent signals from different classes. While the structure of the model $\mathcal{M}(\Theta)$ for different classes may be different in this framework, the parameters in any given class model are still estimated by maximizing the likelihood of a set of training features from that class. Furthermore, the class-conditional likelihood $\mathcal{L}\{\mathbf{Z}|\mathcal{M}(\Theta)\}$ is still used as a measure of class membership, although a "change-of-measure" operation may be required before it can be compared to class-membership measures for other classes. In both the conventional and class-specific scenarios, then, the fundamental problems in stochastic modeling for classification are model parameter estimation and likelihood evaluation.

One possible choice for $\mathcal{M}(\Theta)$ is the linear Gaussian state-space model, or Kalman-filter model, which is general enough to accommodate a wide range of classes and features. This model is supported by an extensive body of theory that has evolved out of the tracking, control, and optimal filtering arenas [3–6], but it has not traditionally

been used as a classification tool. Another potential choice for $\mathcal{M}(\Theta)$ is the family of HMMs, which have been used very successfully to classify sequences of Fourier- and cepstrum-based features in speech recognition [7, 8]. Traditional HMMs and the popular Baum-Welch training algorithm [9], however, constrain the state vectors to reside in a discrete state space, making them unsuitable for signals whose features vary continuously as a function of time. A component of the present work is the development of continuous-state HMMs (CS-HMMs) that are better able to handle continuously varying feature sequences. The CS-HMM algorithms are then specialized to the linear Gaussian models, which are referred to in this context as hidden Gauss-Markov models (HGMMs).

While the focus in this report is on classification, the theory may also be of interest in tracking applications due to the prominence of the Kalman filter as a tracking tool. A common criticism of the Kalman filter is that it gives too much credence to the model and not enough to the observed data. A symptom of this "narcissistic-model syndrome" is that the error covariances depend only on the *a priori* values of the model parameters and are independent of the observed data. An improved tracker might be obtained by including a parameter-updating scheme during the processing of observed data, as is done during the training phase for an HGMM. Furthermore, the general development of CS-HMMs could provide a framework for investigating tracking algorithms based on non-Gaussian models.

1.2 RELATED LITERATURE

Discrete-state HMMs (DS-HMMs) were developed by Baum and his colleagues in the late 1960s and early 1970s [9–12]. Baum's work includes a method for estimating the sequence of individually most likely hidden states (referred to as the Baum or forward-backward algorithm) and a method for estimating the parameters in the HMM (referred to as the Baum-Welch algorithm). An alternative approach to state estimation is to seek the single most likely sequence of states, which is obtained using the Viterbi algorithm [13, 14]. Dempster *et al.* [15] observed that the Baum-Welch algorithm for estimating the HMM parameters is an example of the EM algorithm. Heiser has since noted that the EM algorithm is a special case of iterative majorization [16].

In 1980, Ferguson [17] helped to solidify HMM theory for speech recognition by outlining the three fundamental problems, namely, state estimation, likelihood evaluation, and parameter estimation. This "HMM paradigm" for speech recognition was developed further by Levinson *et al.* [18].

Most extensions of the basic HMM structure have focused on obtaining more general output densities. Liporace [19] treated HMMs with elliptically symmetric continuous output densities. Juang *et al.* [20] relaxed the elliptical symmetry requirement by treating models with Gaussian-mixture output densities. Other continuous-output HMMs include those whose measurements are governed by an autoregressive process [21–23], a polynomial regression function [24], and a linear Gaussian model [25]. When employed with models having variable-duration states [26], such continuous-output distributions lead to the versatile class of segmental models [27].

In contrast to these continuous-output models, the general class of CS-HMMs has received very little attention in the signal processing literature. Linear Gaussian models, on the other hand, have been extensively studied in the Kalman-filter context, although it is not generally known that these models are examples of CS-HMMs. An excellent review of the early history of Gaussian models in a linear-filtering context was given by Kailath [28]. Methods for estimating parameters in these models are reviewed below.

Most early applications of Kalman-filter models in the engineering literature considered the model parameters (i.e., the transition, output, and covariance matrices) to be known from physical considerations, so that parameter estimation was not an issue. Two notable exceptions are found in the works by Kashyap [29] and Gupta and Mehra [30], who used gradient-based nonlinear optimization techniques to maximize the likelihood as expressed in terms of the measurement innovations. Parameter estimation in Gaussian models is more prominent in the time series and econometrics literature, where the first applications of the EM algorithm for this purpose appeared. In the early 1980s, Shumway and Stoffer [31] and Watson and Engle [32] independently derived EM algorithms for estimating parameters in time-invariant Gaussian models with linear constraints. While the general thrusts of these references are the same, they have a number of distinguishing features. First, while Shumway and Stoffer assume that the output matrix is known *a priori*, Watson and Engle estimate the output matrix along with the other model parameters. Second, Shumway and Stoffer provide an explicit recursive expression for the cross covariance between time-adjacent states, which must be calculated as part of the expectation portion (or E-step) of the EM algorithm. Watson and Engle, on the other hand, obtain the cross covariance implicitly by using an augmented state vector during the Kalman smoothing algorithm that lies at the heart of the E-step. Finally, Shumway and Stoffer derive the EM algorithm by defining and maximizing the auxiliary function explicitly, while Watson and Engle formulate the algorithm in terms of sufficient statistics.

Building on this earlier work, signal processing researchers began using the EM algorithm with linear Gaussian models in the early 1990s. Ziskind and Hertz [33] used this approach to estimate directions of arrival for narrowband autoregressive processes received on a multisensor array. Weinstein *et al.* [34] estimated the parameters in a linear Gaussian model while performing noise removal in signals received on a pair of sensors with known coupling. Digalakis *et al.* [25] extended the EM algorithm to treat time-varying models and used the linear Gaussian model to represent segments of speech. Deng and Shen [35] later provided a decomposition algorithm to speed the processing of the EM algorithm for high-dimensional state spaces. Finally, Elliot and Krishnamurthy [36] derived an efficient filter-based implementation of the correlation matrices required during the E-step of the EM algorithm. This filter-based approach dispenses with the need for the more computationally demanding Kalman smoother, which could make the EM algorithm much more amenable to real-time processing.

An extension of the linear Gaussian model that is particularly relevant to the current work was provided by Sorenson and Alspach [37] and Lo [38], who derived Kalman-filter recursions for models with a Gaussian-mixture prior for the initial state, as opposed to the single Gaussian density that traditionally governs the initial state.

Finally, the unification of HMMs and Kalman filters was initiated a decade ago by Streit [39], who demonstrated that the Baum and Viterbi algorithms for a CS-HMM with Gaussian model densities generate the same state vectors and covariance matrices as are obtained using a fixed-interval Kalman smoother. Streit was considering these algorithms in the context of the tracking and not the classification problem, however, so that likelihood evaluation and model parameter estimation were not at issue. This limited focus allowed Streit to ignore both the scaling constants on the Baum probability densities and the joint densities of adjacent states in his characterization of the optimal state sequences.

1.3 ORIGINAL CONTRIBUTIONS

This report presents a general theory for CS-HMMs, independent of the particular form of the model densities, addressing the state estimation, likelihood evaluation, and parameter estimation problems as outlined for DS-HMMs by Ferguson [17]. While continuous-state versions of the Baum and Viterbi algorithms are, for the most part, straightforward extensions of their discrete-state counterparts and are obtained using methods outlined by Jazwinski [40], the Baum and Viterbi algorithms treat only state estimation and likelihood evaluation. The present research also addresses parameter

estimation, to the extent possible, by giving a general formulation of the EM auxiliary function. This formulation solves the E-step of the EM algorithm. The M-step can then be defined by maximizing the auxiliary function after the form is specified for the model densities.

The CS-HMM results are then specialized to HGMMs. With regard to the Baum and Viterbi algorithms for HGMMs, this research extends the results given by Streit [39] in both scope and substance. Streit demonstrated the equivalence of the Baum and Viterbi state sequences with those generated by the fixed-interval smoothing algorithm developed by Rauch, Tung, and Striebel (commonly referred to as the *RTS smoother*) [41]. It is shown here that Baum's forward-backward algorithm is more naturally equated with the two-filter implementation of the smoother given by Mayne [42] and Fraser and Potter [43], which employs forward- and backward-running Kalman filters and then estimates the state sequence by optimally combining the estimates from the two filters. The Viterbi algorithm does lead naturally to the state-estimation portion of the RTS smoother, but, as noted by Streit [39], it does not give the state error covariances. The covariance matrices from the RTS algorithm are obtained here using a new joint-density marginalization approach, which provides a more natural derivation of the RTS algorithm from the HMM point of view.

This report presents a new *Gaussian refactorization lemma*. While the original motivation for this lemma was to characterize a recurrent set of operations when deriving the HGMM Baum and Viterbi algorithms, it turns out that this lemma provides a natural alternative derivation of the Kalman filter recursions.

The conditional joint density of time-adjacent states in HGMMs is also derived. In addition to facilitating the theoretical development of the parameter estimation algorithm, this derivation leads to a new expression for the cross covariance between states that is considerably simpler than the recursive definition given by Shumway and Stoffer [31]. The simplicity of the new expression occurs because the cross-covariance matrix is viewed within the larger context of the joint density, as opposed to its being derived by taking the expectation of the appropriate outer product.

In addressing likelihood evaluation and parameter estimation for HGMMs, this report shows two additional equivalences between HMMs and Kalman-filter models. First, Baum's forward recursion for marginalizing out the states from the joint distribution of the states and measurements yields the classical expression for the measurement likelihood from Kalman filter theory [44]. Second, the continuous-state formulation of the Baum-Welch auxiliary function leads to the existing EM algorithm for estimating the parameters in Kalman-filter models.

In addition to these results, the CS-HMM and HGMM algorithms are extended to accommodate prior state densities that are composed of mixtures. These new developments provide substantial extensions of previous work with mixture-based Kalman filters [37, 38], first by generalizing the mixture-based algorithms to the larger class of CS-HMMs, and then by addressing the smoothing and parameter estimation problems. In the classification context, the inclusion of mixture-based prior densities allows the model to accommodate even greater amounts of within-class variability than can be handled by "single-mode" models.

All of the above results are given for models whose parameters are time varying. For HGMMs with parameters that are constant across time, the measurement likelihood is shown to be invariant to a family of similarity transformations of the model parameters.

This report provides a unified context from which to view HMMs and Kalman filters for both classification and tracking. The literature review given above provides a sampling of the very rich history that has unfolded in connection with these families of tools. For the most part, this history has evolved along two separate paths, one leading from Baum, the other leading from Kalman. It is shown here that these two paths are really one.

2. CONTINUOUS-STATE HIDDEN MARKOV MODELS

In general, discrete-time HMMs represent a sequence of observed M-dimensional measurements $\mathbf{Z}_N = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$, made at times t_n, $n = 1, 2, \ldots, N$, as probabilistic functions of a sequence of unobservable L-dimensional states $\mathbf{X}_N = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_0$ is governed by a prior distribution and does not correspond to a measurement. If the $\mathbf{x}_n$ are constrained to take values from a discrete finite-sized alphabet, then the model is a DS-HMM. In such models, there is no need to specify the actual element values for vector $\mathbf{x}_n$. Only the index into the alphabet containing the different values of $\mathbf{x}_n$ need be provided. This index is usually denoted as i or j, and the event $q_n(i)$ indicates that the ith state has occurred at time t_n. If the elements of $\mathbf{x}_n$ are free to assume values on the real line, then the model is a CS-HMM. In these models, the element values must be specified since the indexing of all possible state vectors would require an uncountable number of index values.

Regardless of whether the states are discrete or continuous, they are usually assumed to obey a first-order Markov process, where each state depends only on the state that occurred at the previous time instant. These first-order HMMs are characterized by three model probability functions. First, the state-transition distribution governs the probability of moving from one state to another in a single time step. Second, the output distribution governs the probability of obtaining a particular measurement, given the value of the state vector. Finally, the prior state distribution governs the probability that the state at time t_0 will take on a particular value.

For DS-HMMs, the prior state probabilities and state-transition probabilities are discrete sets of numbers denoted by π_i and a_{ij}, respectively, for states with indices i and j. The output distribution can be either discrete or continuous, and is denoted $b_j(\mathbf{z}_n)$. While discrete-state models work quite well for certain classes of signals and features (e.g., spectral or cepstral coefficients of speech), they are ill-suited for feature sequences that follow a continuous trajectory through state space (e.g., the instantaneous frequency of a wandering tonal). For these types of signals, continuous-state models allow a more accurate representation. The model distributions are denoted in this case by the density functions $p(\mathbf{x}_0|\theta_0)$, $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \theta_X)$, and $p(\mathbf{z}_n|\mathbf{x}_n, \theta_Z)$. The structure of these model densities is usually known (e.g., Gaussian, Rayleigh, gamma), but the parameter sets θ_0, θ_X, and θ_Z must be estimated for each class from training data. For convenience in notation, only those probability models having well-defined density functions are considered here.

The notation for discrete- and continuous-state models is summarized in table 1. It is important to note that, when moving from discrete to continuous models, expressions such as $p(\mathbf{x};\theta)$ no longer designate probabilities, but likelihoods. While it would be technically more accurate to use expressions such as $\mathcal{L}(\mathbf{x};\theta)$ for CS-HMMs, the p-notation is used to follow conventions established in existing literature. The notation is abused even further by using the same symbol for the likelihood of an event $\mathbf{x}$ (i.e., a particular numerical value or sequence of values) and the probability density function of a random variable $\mathbf{x}$. The context can serve as a guide for the meaning of the expression, however, since the difference between the likelihood function and the probability density function is merely a matter of whether θ or $\mathbf{x}$ is treated as unknown, and the likelihood is just the likelihood function evaluated at a particular value.

Table 1. Notation for Hidden Markov Models

	Discrete-State Model	Continuous-State Model
State	i, j	$\mathbf{x}_n$
Measurement	$\mathbf{z}_n$	$\mathbf{z}_n$
State-Transition Probability	a_{ij}	$p(\mathbf{x}_n\|\mathbf{x}_{n-1},\theta_X)$
Output Probability	$b_j(\mathbf{z}_n)$	$p(\mathbf{z}_n\|\mathbf{x}_n,\theta_Z)$
Prior State Probability	π_i	$p(\mathbf{x}_0\|\theta_0)$

This section derives algorithms for likelihood evaluation, state estimation, and model parameter estimation with CS-HMMs. For discrete-state models, these problems are solved using the Baum, Viterbi, and Baum-Welch algorithms [9, 13, 17, 18]. The continuous-state versions of these algorithms are derived here. While the continuous-state analog to the Baum-Welch algorithm cannot be completely specified without assuming explicit forms for the model densities, the auxiliary function for the EM algorithm can be formulated in a density-independent fashion. In addition to these developments, the CS-HMM and associated algorithms are extended to include prior state distributions consisting of a mixture of densities.

2.1 BAUM PROBABILITY DENSITIES

The state evolution in CS-HMMs is characterized by the joint density of the measurement and state sequences, which is given by

$$p(\mathbf{Z}_N, \mathbf{X}_N) = p(\mathbf{x}_0) \prod_{n=1}^{N} p(\mathbf{z}_n|\mathbf{x}_n)\, p(\mathbf{x}_n|\mathbf{x}_{n-1}), \tag{1}$$

where the explicit notational dependence on the model parameters has been dropped. This expression makes two basic assumptions: (1) the states follow a first-order Markov model and (2) the measurements are conditionally independent, given the states. Class assignments are made in signal classification using the measurement likelihood, which is obtained by marginalizing equation (1) over all possible state sequences, giving

$$p(\mathbf{Z}_N) = \int d\mathbf{X}_N\, p(\mathbf{Z}_N, \mathbf{X}_N). \tag{2}$$

Here, the shorthand $\int d\mathbf{X}_N$ denotes the multiple integral $\int d\mathbf{x}_0 \cdots \int d\mathbf{x}_N$, where each single integral $\int d\mathbf{x}_n$ is an L-dimensional integration over state space. For the discussion below, it is also necessary to introduce the partial measurement sequence $\mathbf{Z}_n = \left\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\right\}$ and its complement in $\mathbf{Z}_N$, denoted $\mathbf{Z}_n^{\mathrm{C}} = \left\{\mathbf{z}_{n+1}, \mathbf{z}_{n+2}, \ldots, \mathbf{z}_N\right\}$.

Because of the intractable computational load required to evaluate the discrete equivalent of equation (2), Baum *et al.* [9] developed recursive functions to characterize and marginalize the joint probability for DS-HMMs. These functions are defined in table 2 using both the discrete-state and the continuous-state notation. The algorithm for computing these probability (or probability-density) functions is referred to as the Baum, or forward-backward, algorithm. The Baum recursions for DS-HMMs are given in table 3 for comparison with the continuous-state recursions derived in this section.

2.1.1 Forward Densities

The forward densities are defined as

$$\begin{aligned}
\alpha(\mathbf{x}_n) &= p(\mathbf{Z}_n, \mathbf{x}_n) \\
&= p(\mathbf{z}_n|\mathbf{Z}_{n-1}, \mathbf{x}_n)\, p(\mathbf{Z}_{n-1}, \mathbf{x}_n) \\
&= p(\mathbf{z}_n|\mathbf{x}_n) \int d\mathbf{x}_{n-1}\, p(\mathbf{Z}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1}) \\
&= p(\mathbf{z}_n|\mathbf{x}_n) \int d\mathbf{x}_{n-1}\, p(\mathbf{x}_n|\mathbf{Z}_{n-1}, \mathbf{x}_{n-1})\, p(\mathbf{Z}_{n-1}, \mathbf{x}_{n-1}) \\
&= p(\mathbf{z}_n|\mathbf{x}_n) \int d\mathbf{x}_{n-1}\, p(\mathbf{x}_n|\mathbf{x}_{n-1})\ \alpha(\mathbf{x}_{n-1}),
\end{aligned} \tag{3}$$

where the recursion defined in this last expression is initialized as $\alpha(\mathbf{x}_0) = p(\mathbf{x}_0)$.

Table 2. Baum Probability Functions

	Discrete-State Definition	Continuous-State Definition
Forward Probability	$\alpha_n(i) = p\{q_n(i), \mathbf{Z}_n\}$	$\alpha(\mathbf{x}_n) = p(\mathbf{x}_n, \mathbf{Z}_n)$
Backward Probability	$\beta_n(i) = p\{\mathbf{Z}_n^{\mathsf{C}} \mid q_n(i)\}$	$\beta(\mathbf{x}_n) = p(\mathbf{Z}_n^{\mathsf{C}} \mid \mathbf{x}_n)$
Conditional State Probability	$\gamma_n(i) = p\{q_n(i) \mid \mathbf{Z}_N\}$	$\gamma(\mathbf{x}_n) = p(\mathbf{x}_n \mid \mathbf{Z}_N)$
Conditional Joint State Probability	$\gamma_n(i,j) = p\{q_{n-1}(i), q_n(j) \mid \mathbf{Z}_N\}$	$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) = p(\mathbf{x}_n, \mathbf{x}_{n-1} \mid \mathbf{Z}_N)$

This recursion can be alternatively defined by

$$\alpha(\mathbf{x}_n) = p(\mathbf{z}_n|\mathbf{x}_n) \int d\mathbf{x}_{n-1}\, \delta(\mathbf{x}_n, \mathbf{x}_{n-1}), \tag{4}$$

where $\delta(\mathbf{x}_n, \mathbf{x}_{n-1})$ is defined as

$$\begin{aligned} \delta(\mathbf{x}_n, \mathbf{x}_{n-1}) &= p(\mathbf{Z}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1}) \\ &= p(\mathbf{x}_n|\mathbf{Z}_{n-1}, \mathbf{x}_{n-1})\, p(\mathbf{Z}_{n-1}, \mathbf{x}_{n-1}) \\ &= p(\mathbf{x}_n|\mathbf{x}_{n-1})\, \alpha(\mathbf{x}_{n-1}). \end{aligned} \tag{5}$$

The recursion for $\alpha(\mathbf{x}_n)$ can therefore be performed by first computing and marginalizing $\delta(\mathbf{x}_n, \mathbf{x}_{n-1})$ and then multiplying by the output density.

2.1.2 Backward Densities

The backward densities are needed primarily as an intermediate step in calculating the conditional densities $\gamma(\mathbf{x}_n)$ and $\gamma(\mathbf{x}_{n+1}, \mathbf{x}_n)$. The recursion for the backward densities is given by

$$\begin{aligned} \beta(\mathbf{x}_{n-1}) &= p(\mathbf{Z}_{n-1}^{\mathsf{C}}|\mathbf{x}_{n-1}) \\ &= \frac{1}{p(\mathbf{x}_{n-1})} \int d\mathbf{x}_n\, p(\mathbf{z}_n, \mathbf{Z}_n^{\mathsf{C}}, \mathbf{x}_n, \mathbf{x}_{n-1}) \\ &= \frac{1}{p(\mathbf{x}_{n-1})} \int d\mathbf{x}_n\, p(\mathbf{z}_n|\mathbf{x}_n)\, p(\mathbf{Z}_n^{\mathsf{C}}|\mathbf{x}_n, \mathbf{x}_{n-1})\, p(\mathbf{x}_n, \mathbf{x}_{n-1}) \\ &= \int d\mathbf{x}_n\, p(\mathbf{z}_n|\mathbf{x}_n)\, p(\mathbf{Z}_n^{\mathsf{C}}|\mathbf{x}_n)\, p(\mathbf{x}_n|\mathbf{x}_{n-1}) \\ &= \int d\mathbf{x}_n\, p(\mathbf{z}_n|\mathbf{x}_n)\, p(\mathbf{x}_n|\mathbf{x}_{n-1})\, \beta(\mathbf{x}_n). \end{aligned} \tag{6}$$

Table 3. Baum Algorithm for DS-HMMs

Forward Probability (Initialization)	$\alpha_0(i) = \pi_i$ for all i
Forward Probability (Recursion)	$\alpha_n(j) = b_j(z_n) \sum_i a_{ij}\, \alpha_{n-1}(i)$
Measurement Probability	$p(\mathbf{Z}_N) = \sum_i \alpha_N(i)$
Backward Probability (Initialization)	$\beta_N(j) = 1$ for all j
Backward Probability (Recursion)	$\beta_n(i) = \sum_j a_{ij}\, b_j(z_{n+1})\, \beta_{n+1}(j)$
Conditional State Probability	$\gamma_n(i) = \frac{1}{p(\mathbf{Z}_N)}\, \alpha_n(i)\, \beta_n(i)$
Conditional Joint State Probability	$\gamma_n(i,j) = \frac{1}{p(\mathbf{Z}_N)}\, a_{ij}\, \alpha_{n-1}(i)\, b_j(z_n)\, \beta_n(j)$

In this expression, $\beta(\mathbf{x}_n)$ is undefined at the terminal time t_N because $\mathbf{Z}_N^{\mathrm{C}}$ is empty. The DS-HMM literature usually defines $\beta_N(j) = 1$ for all j. In the continuous-state case, such a definition is problematic because $\beta(\mathbf{x}_N) = 1$ is not an integrable probability density. To be formally correct, the recursion should be started at time t_{N-1}, and $\beta(\mathbf{x}_{N-1}) = \int d\mathbf{x}_N\, p(\mathbf{z}_N|\mathbf{x}_N)\, p(\mathbf{x}_N|\mathbf{x}_{N-1})$ should be defined as a special case. This approach is notationally inconvenient, however, because other densities are defined as products of α and β, which would cause a proliferation of special cases. So, as a purely notational mechanism, $\beta(\mathbf{x}_N)$ is set to unity for all $\mathbf{x}_N$.

The backward density can be alternatively defined as

$$\beta(\mathbf{x}_{n-1}) = \int d\mathbf{x}_n\, p(\mathbf{x}_n|\mathbf{x}_{n-1})\, \psi(\mathbf{x}_n), \tag{7}$$

where

$$\begin{aligned} \psi(\mathbf{x}_n) &= p(\mathbf{Z}_{n-1}^{\mathrm{C}}|\mathbf{x}_n) \\ &= p(\mathbf{z}_n|\mathbf{Z}_n^{\mathrm{C}}, \mathbf{x}_n)\, p(\mathbf{Z}_n^{\mathrm{C}}|\mathbf{x}_n) \\ &= p(\mathbf{z}_n|\mathbf{x}_n)\, \beta(\mathbf{x}_n). \end{aligned} \tag{8}$$

The backward recursion thus proceeds by first computing $\psi(\mathbf{x}_n)$, then multiplying by the transition density, and finally marginalizing over $\mathbf{x}_n$.

2.1.3 *Conditional State Densities*

The conditional state probability densities characterize the stochastic properties of individual states when conditioned on the observed measurements. These densities can be maximized to determine the sequence of individually most likely states. They are also important for parameter estimation. These densities are defined as

$$\begin{aligned}
\gamma(\mathbf{x}_n) &= p(\mathbf{x}_n|\mathbf{Z}_N) \\
&= \frac{1}{p(\mathbf{Z}_N)}\, p(\mathbf{Z}_n^{\mathrm{C}}, \mathbf{Z}_n, \mathbf{x}_n) \\
&= \frac{1}{p(\mathbf{Z}_N)}\, p(\mathbf{Z}_n^{\mathrm{C}}|\mathbf{Z}_n, \mathbf{x}_n)\, p(\mathbf{Z}_n, \mathbf{x}_n) \\
&= \frac{1}{p(\mathbf{Z}_N)}\, p(\mathbf{Z}_n^{\mathrm{C}}|\mathbf{x}_n)\, p(\mathbf{Z}_n, \mathbf{x}_n) \\
&= \frac{1}{p(\mathbf{Z}_N)}\, \beta(\mathbf{x}_n)\, \alpha(\mathbf{x}_n). \qquad (9)
\end{aligned}$$

2.1.4 *Conditional Joint State Densities*

Finally, the conditional joint state densities characterize the relational properties of time-adjacent states when conditioned on the measurements. They are defined as

$$\begin{aligned}
\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) &= p(\mathbf{x}_n, \mathbf{x}_{n-1}|\mathbf{Z}_N) \\
&= \frac{1}{p(\mathbf{Z}_N)}\, p(\mathbf{Z}_{n-1}, \mathbf{Z}_{n-1}^{\mathrm{C}}, \mathbf{x}_n, \mathbf{x}_{n-1}) \\
&= \frac{1}{p(\mathbf{Z}_N)}\, p(\mathbf{Z}_{n-1}^{\mathrm{C}}|\mathbf{Z}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1})\, p(\mathbf{Z}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1}) \\
&= \frac{1}{p(\mathbf{Z}_N)}\, p(\mathbf{Z}_{n-1}^{\mathrm{C}}|\mathbf{x}_n)\, p(\mathbf{Z}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1}) \\
&= \frac{1}{p(\mathbf{Z}_N)}\, \psi(\mathbf{x}_n)\, \delta(\mathbf{x}_n, \mathbf{x}_{n-1}). \qquad (10)
\end{aligned}$$

This expression looks quite different from the discrete-state version because it has been derived directly in terms of $\psi(\mathbf{x}_n)$ and $\delta(\mathbf{x}_n, \mathbf{x}_{n-1})$. Substituting the definitions of $\psi(\mathbf{x}_n)$ and $\delta(\mathbf{x}_n, \mathbf{x}_{n-1})$ in terms of $\beta(\mathbf{x}_n)$ and $\alpha(\mathbf{x}_{n-1})$ provides the more familiar-looking expression

$$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) = \frac{1}{p(\mathbf{Z}_N)}\, p(\mathbf{z}_n|\mathbf{x}_n)\, \beta(\mathbf{x}_n)\, p(\mathbf{x}_n|\mathbf{x}_{n-1})\, \alpha(\mathbf{x}_{n-1}). \qquad (11)$$

2.2 LIKELIHOOD EVALUATION

The measurement likelihood can be expressed in terms of the forward and backward densities by writing

$$\begin{aligned} p(\mathbf{Z}_N) &= \int d\mathbf{x}_n \, p(\mathbf{Z}_N, \mathbf{x}_n) \\ &= \int d\mathbf{x}_n \, p(\mathbf{Z}_n^{\mathsf{C}} | \mathbf{Z}_n, \mathbf{x}_n) \, p(\mathbf{Z}_n, \mathbf{x}_n) \\ &= \int d\mathbf{x}_n \, p(\mathbf{Z}_n^{\mathsf{C}} | \mathbf{x}_n) \, p(\mathbf{Z}_n, \mathbf{x}_n) \\ &= \int d\mathbf{x}_n \, \beta(\mathbf{x}_n) \, \alpha(\mathbf{x}_n). \end{aligned} \tag{12}$$

This result demonstrates that the definition given for the conditional state density is properly normalized; that is, $\int d\mathbf{x}_n \gamma(\mathbf{x}_n) = 1$ for all n. A simpler expression is obtained by evaluating the likelihood at t_N, where it is

$$p(\mathbf{Z}_N) = \int d\mathbf{x}_N \, p(\mathbf{Z}_N, \mathbf{x}_N) = \int d\mathbf{x}_N \, \alpha(\mathbf{x}_N). \tag{13}$$

This outcome is consistent with equation (12) since $\beta(\mathbf{x}_N) = 1$. The Baum algorithm and likelihood evaluation formulas for CS-HMMs are collected in table 4.

2.3 VITERBI ALGORITHM

The Viterbi algorithm is a two-pass dynamic programming algorithm [45] that evaluates the maximum *a posteriori* (MAP) estimate of the state sequence, defined as

$$\hat{\mathbf{X}}_N = \arg\max_{\mathbf{X}_N} p(\mathbf{X}_N | \mathbf{Z}_N) = \arg\max_{\mathbf{X}_N} p(\mathbf{X}_N, \mathbf{Z}_N), \tag{14}$$

where $p(\mathbf{Z}_N)$ is constant for any $\mathbf{Z}_N$ and does not affect the arg max operation. The forward pass propagates a function $\phi(\mathbf{x}_n)$, which is initialized as $\phi(\mathbf{x}_0) = p(\mathbf{x}_0)$. The forward recursion is then defined for $n = 1, \ldots, N$ as

$$\phi(\mathbf{x}_n) = p(\mathbf{z}_n | \mathbf{x}_n) \max_{\mathbf{x}_{n-1}} \left\{ p(\mathbf{x}_n | \mathbf{x}_{n-1}) \, \phi(\mathbf{x}_{n-1}) \right\}. \tag{15}$$

This expression is similar to Baum's forward density $\alpha(\mathbf{x}_n)$, except that it contains a maximization instead of a marginalization.

The backward pass of the Viterbi algorithm is a back substitution operation. The optimal estimate at time t_N is determined as

$$\hat{\mathbf{x}}_N = \arg\max_{\mathbf{x}_N} \phi(\mathbf{x}_N), \tag{16}$$

which is then used to initialize the backward recursions, defined as

$$\hat{\mathbf{x}}_{n-1} = \arg\max_{\mathbf{x}_{n-1}} \left\{ p(\hat{\mathbf{x}}_n | \mathbf{x}_{n-1}) \, \phi(\mathbf{x}_{n-1}) \right\}. \tag{17}$$

Table 4. Baum Algorithm for CS-HMMs

Forward Density (Initialization)	$\alpha(\mathbf{x}_0) = p(\mathbf{x}_0)$
Forward Density (Stage 1 Recursion)	$\delta(\mathbf{x}_n, \mathbf{x}_{n-1}) = p(\mathbf{x}_n \vert \mathbf{x}_{n-1})\, \alpha(\mathbf{x}_{n-1})$
Forward Density (Stage 2 Recursion)	$\alpha(\mathbf{x}_n) = p(\mathbf{z}_n \vert \mathbf{x}_n) \int d\mathbf{x}_{n-1}\, \delta(\mathbf{x}_n, \mathbf{x}_{n-1})$
Measurement Likelihood	$p(\mathbf{Z}_N) = \int d\mathbf{x}_n\, \alpha(\mathbf{x}_N)$
Backward Density (Initialization)	$\beta(\mathbf{x}_N) = 1$ for all $\mathbf{x}_N$
Backward Density (Stage 1 Recursion)	$\psi(\mathbf{x}_n) = p(\mathbf{z}_n \vert \mathbf{x}_n)\, \beta(\mathbf{x}_n)$
Backward Density (Stage 2 Recursion)	$\beta(\mathbf{x}_{n-1}) = \int d\mathbf{x}_n\, p(\mathbf{x}_n \vert \mathbf{x}_{n-1})\, \psi(\mathbf{x}_n)$
Conditional State Density	$\gamma(\mathbf{x}_n) = \frac{1}{p(\mathbf{Z}_N)}\, \alpha(\mathbf{x}_n)\, \beta(\mathbf{x}_n)$
Conditional Joint State Density	$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) = \frac{1}{p(\mathbf{Z}_N)}\, \psi(\mathbf{x}_n)\, \delta(\mathbf{x}_n, \mathbf{x}_{n-1})$

2.4 PARAMETER ESTIMATION

Hidden-state models are natural candidates for the EM algorithm, which distinguishes three types of data: the *incomplete data*, which in this context are the observed measurements; the *missing data*, which are the hidden states; and the *complete* data, which are the concatenation of the incomplete and missing data. Since the joint density in equation (1) is the likelihood of the complete data, that function is referred to here as the *complete-data likelihood function* (CDLF).

For time-varying models, parameter estimation using the EM algorithm requires the use of a multiple-sequence training set since no time averaging can be performed and since EM parameter estimation with a single piece of data merely repeats the previous estimate at each iteration. Single-sequence training can be performed for time-invariant models since the parameters can be averaged over time, although classification models so obtained will likely have poor generalization performance.

The multiple-measurement training set is denoted as

$$\mathcal{Z} = \left\{ \mathbf{Z}_{N_1}^1, \mathbf{Z}_{N_2}^2, \ldots, \mathbf{Z}_{N_K}^K \right\}, \tag{18}$$

where $\mathbf{Z}_{N_k}^k = \left\{ \mathbf{z}_1^k, \mathbf{z}_2^k, \ldots, \mathbf{z}_{N_k}^k \right\}$ is the kth training sequence. The lengths of these training sequences are not constrained to be equal, although the sequences are assumed to be arranged so that $N_1 \geq N_2 \geq \cdots N_K$. Even with multiple training sequences, however, a difficulty arises due to the unequal lengths of the training sequences. Parameters in the densities $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \Theta)$ and $p(\mathbf{z}_n|\mathbf{x}_n, \Theta)$ corresponding to large n are estimated from fewer and fewer training sequences as n successively exceeds the lengths of the shorter measurements. If there is a unique longest training sequence, then parameters corresponding to time samples that occur after the end of the second longest sequence are "estimated" from a single measurement. This problem might be addressed by truncating the longer measurement sequences to some predetermined value or by using some type of time-warping to obtain equal-length measurements, depending on the application. In what follows, the above difficultly is assumed to have been dealt with in the appropriate manner.

Notationally, the unequal sequence lengths are accommodated by introducing the variable K_n, which is, for each time t_n, the number of training sequences whose length equals or exceeds n. This variable represents the effective number of training samples available at t_n. Recalling that the measurements are assumed to be arranged in decreasing order of length and defining $N_{\max} = \max(N_k)$, then for any function $f(\mathbf{x}_n^k)$,

$$\sum_{k=1}^{K} \sum_{n=1}^{N_k} f(\mathbf{x}_n^k) = \sum_{n=1}^{N_{\max}} \sum_{k=1}^{K_n} f(\mathbf{x}_n^k). \tag{19}$$

The EM algorithm generates parameter estimates iteratively, where at each iteration the estimates are chosen to maximize the conditional expectation of the CDLF, given the observed data and the model parameters from the previous algorithm iteration. Denoting by $\mathcal{X} = \left\{ \mathbf{X}_{N_k}^k, k = 1, \ldots, K \right\}$ the collection of state sequences corresponding to the measurements in the training set, the CDLF for the training set is

$$\begin{aligned} p(\mathcal{Z}, \mathcal{X}|\Theta) &= \prod_{k=1}^{K} p(\mathbf{Z}_{N_k}^k, \mathbf{X}_{N_k}^k|\Theta) \\ &= \prod_{k=1}^{K} p(\mathbf{x}_0^k|\theta_0) \prod_{n=1}^{N_k} p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k, \theta_X)\, p(\mathbf{z}_n^k|\mathbf{x}_n^k, \theta_Z). \end{aligned} \tag{20}$$

Each iteration of the EM algorithm generates estimates

$$\Theta^{i+1} = \arg\max_{\Theta} Q(\Theta, \Theta^i) \tag{21}$$

that maximize the auxiliary function

$$\begin{aligned} Q(\Theta, \Theta^i) &= E_{\mathcal{X}|\mathcal{Z},\Theta^i}\Big\{\log p(\mathcal{Z}, \mathcal{X}|\Theta)\Big\} \\ &= \int d\mathcal{X}\, p(\mathcal{X}|\mathcal{Z}, \Theta^i) \log p(\mathcal{Z}, \mathcal{X}|\Theta) \\ &= \prod_{\ell=1}^{K} \int d\mathbf{X}_{N_\ell}^{\ell}\, p(\mathbf{X}_{N_\ell}^{\ell}|\mathbf{Z}_{N_\ell}^{\ell}, \Theta^i) \log p(\mathcal{Z}, \mathcal{X}|\Theta). \end{aligned} \tag{22}$$

The E-step evaluates the auxiliary function at Θ^i from the previous iteration, yielding a function of the unknown parameters only. The M-step then generates optimal estimates of the unknown parameters by maximizing the auxiliary function. Since the M-step requires differentiation with respect to the model parameters, it cannot be specified without imposing a particular form for the model densities. The E-step can be specified in greater detail, however, by imposing the structure of the HMM.

Since the CDLF consists of a product of model densities that are parameterized by separate subsets of the model parameters, the auxiliary function is decomposed as

$$Q(\Theta, \Theta^i) = Q_0(\theta_0, \Theta^i) + Q_X(\theta_X, \Theta^i) + Q_Z(\theta_Z, \Theta^i), \tag{23}$$

where each component corresponds to one of the three model densities. The parameters in the model's initial-state density are estimated by maximizing the component

$$\begin{aligned} Q_0(\theta_0, \Theta^i) &= E_{\mathcal{X}|\mathcal{Z},\Theta^i}\left\{\log\left[\prod_{k=1}^{K} p(\mathbf{x}_0^k|\theta_0)\right]\right\} \\ &= \prod_{\ell=1}^{K} \int d\mathbf{X}_{N_\ell}^{\ell}\, p(\mathbf{X}_{N_\ell}^{\ell}|\mathbf{Z}_{N_\ell}^{\ell}, \Theta^i) \left\{\sum_{k=1}^{K} \log p(\mathbf{x}_0^k|\theta_0)\right\} \\ &= \sum_{k=1}^{K} \prod_{\ell=1}^{K} \int d\mathbf{X}_{N_\ell}^{\ell}\, p(\mathbf{X}_{N_\ell}^{\ell}|\mathbf{Z}_{N_\ell}^{\ell}, \Theta^i) \log p(\mathbf{x}_0^k|\theta_0) \\ &= \sum_{k=1}^{K} \int d\mathbf{X}_{N_k}^{k}\, p(\mathbf{X}_{N_k}^{k}|\mathbf{Z}_{N_k}^{k}, \Theta^i) \log p(\mathbf{x}_0^k|\theta_0) \\ &= \sum_{k=1}^{K} \int d\mathbf{x}_0^k\, p(\mathbf{x}_0^k|\mathbf{Z}_{N_k}^{k}, \Theta^i) \log p(\mathbf{x}_0^k|\theta_0) \\ &= \sum_{k=1}^{K} \int d\mathbf{x}_0^k\, \gamma(\mathbf{x}_0^k) \log p(\mathbf{x}_0^k|\theta_0), \end{aligned} \tag{24}$$

where $\gamma(\mathbf{x}_0^k) = p(\mathbf{x}_0^k|\mathbf{Z}_{N_k}^{k}, \Theta^i)$ is the conditional state density for time t_0 under the old parameter values in Θ^i.

The parameters in the model's transition density are obtained during each EM iteration by maximizing the auxiliary-function component

$$
\begin{aligned}
Q_X(\theta_X, \Theta^i) &= E_{\mathcal{X}|\mathcal{Z},\Theta^i}\left\{\log\left[\prod_{k=1}^{K}\prod_{n=1}^{N_k} p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k, \theta_X)\right]\right\} \\
&= \prod_{\ell=1}^{K}\int d\mathbf{X}_{N_\ell}^{\ell}\; p(\mathbf{X}_{N_\ell}^{\ell}|\mathbf{Z}_{N_\ell}^{\ell}, \Theta^i)\left\{\sum_{k=1}^{K}\sum_{n=1}^{N_k}\log p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k, \theta_X)\right\} \\
&= \sum_{k=1}^{K}\int d\mathbf{X}_{N_k}^{k}\; p(\mathbf{X}_{N_k}^{k}|\mathbf{Z}_{N_k}^{k}, \Theta^i)\left\{\sum_{n=1}^{N_k}\log p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k, \theta_X)\right\} \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N_k}\int d\mathbf{X}_{N_k}^{k}\; p(\mathbf{X}_{N_k}^{k}|\mathbf{Z}_{N_k}^{k}, \Theta^i)\log p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k, \theta_X) \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N_k}\iint d\mathbf{x}_n^k\, d\mathbf{x}_{n-1}^k\; p(\mathbf{x}_n^k, \mathbf{x}_{n-1}^k|\mathbf{Z}_{N_k}^{k}, \Theta^i)\log p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k, \theta_X) \\
&= \sum_{n=1}^{N_{\max}}\sum_{k=1}^{K_n}\int d\mathbf{x}_n^k\int d\mathbf{x}_{n-1}^k\;\gamma(\mathbf{x}_n^k, \mathbf{x}_{n-1}^k)\log p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k, \theta_X), \qquad (25)
\end{aligned}
$$

where $\gamma(\mathbf{x}_n^k, \mathbf{x}_{n-1}^k) = p(\mathbf{x}_n^k, \mathbf{x}_{n-1}^k|\mathbf{Z}_{N_k}^{k}, \Theta^i)$ is the conditional joint state density obtained from Baum's forward-backward algorithm with the kth measurement. Note that one of the summations over the members of the measurement training set drops out in the third step of equation (25) because the summand is zero except when $\ell = k$.

Finally, the parameters in the output density are obtained by maximizing the auxiliary-function component

$$
\begin{aligned}
Q_Z(\theta_Z, \Theta^i) &= E_{\mathcal{X}|\mathcal{Z},\Theta^i}\left\{\log\left[\prod_{k=1}^{K}\prod_{n=1}^{N_k} p(\mathbf{z}_n^k|\mathbf{x}_n^k, \theta_Z)\right]\right\} \\
&= \prod_{\ell=1}^{K}\int d\mathbf{X}_{N_\ell}^{\ell}\; p(\mathbf{X}_{N_\ell}^{\ell}|\mathbf{Z}_{N_\ell}^{\ell}, \Theta^i)\left\{\sum_{k=1}^{K}\sum_{n=1}^{N_k} log\, p(\mathbf{z}_n^k|\mathbf{x}_n^k, \theta_Z)\right\} \\
&= \sum_{k=1}^{K}\int d\mathbf{X}_{N_k}^{k}\; p(\mathbf{X}_{N_k}^{k}|\mathbf{Z}_{N_k}^{k}, \Theta^i)\left\{\sum_{n=1}^{N_k}\log p(\mathbf{z}_n^k|\mathbf{x}_n^k, \theta_Z)\right\} \\
&= \sum_{n=1}^{N_{\max}}\sum_{k=1}^{K_n}\int d\mathbf{x}_n^k\;\gamma(\mathbf{x}_n^k)\log p(\mathbf{z}_n^k|\mathbf{x}_n^k, \theta_Z), \qquad (26)
\end{aligned}
$$

where $\gamma(\mathbf{x}_n^k) = p(\mathbf{x}_n^k|\mathbf{Z}_{N_k}^{k}, \Theta^i)$ is the conditional state density obtained from Baum's algorithm. The above results are summarized in table 5.

Table 5. EM Auxiliary-Function Components for CS-HMMs

Transition Density Parameters	$Q_X(\theta_X, \Theta') = \sum_{n=1}^{N_{\max}} \sum_{k=1}^{K_n} \int dx_n^k \int dx_{n-1}^k \, \gamma(x_n^k, x_{n-1}^k) \log p(x_n^k \mid x_{n-1}^k, \theta_X)$
Output Density Parameters	$Q_Z(\theta_Z, \Theta') = \sum_{n=1}^{N_{\max}} \sum_{k=1}^{K_n} \int dx_n^k \, \gamma(x_n^k) \log p(z_n^k \mid x_n^k, \theta_Z)$
Prior State Density Parameters	$Q_0(\theta_0, \Theta') = \sum_{k=1}^{K} \int dx_0^k \, \gamma(x_0^k) \log p(x_0^k \mid \theta_0)$

If the continuous variables in this table are replaced with their discrete counterparts and the auxiliary-function components are maximized over those variables, the Baum-Welch re-estimation formulas [9] are obtained. Estimation of the parameters in the transition density is simplified in this case because the a_{ij} enter into the model linearly, subject to the constraint that the "exiting probabilities" for the ith state must sum to unity. The re-estimation formula for a_{ij} is therefore obtained by solving the constrained optimization problem whose Lagrangian is

$$\tilde{Q}_X = \sum_{n=1}^{N_{\max}} \sum_{k=1}^{K_n} \sum_i \sum_j \gamma_n^k(i,j) \log a_{ij} + \lambda \left(1 - \sum_j a_{ij} \right). \tag{27}$$

A similar Lagrangian is used to obtain the output probabilities for discrete measurement spaces.

2.5 MIXED-MODE MODELS

This section extends the CS-HMM by letting the initial state be governed by the mixture of densities

$$p(x_0) = \sum_{j=1}^{J} \rho_j \, p(x_0 \mid j), \tag{28}$$

where $p(x_0|j)$ is the jth mode in the mixture, j is the mode-assignment index, and $\rho_j = p(j)$ is the mode-assignment probability, or mixing parameter. Since j is a discrete variable, ρ_j is a probability measure and not a density. As usual, the mixture is assumed to be a convex combination with $\rho_j \geq 0$ for all j and $\sum_{j=1}^{J} \rho_j = 1$. For reasons that will later become clear, the resulting model is referred to as a "mixed-mode" CS-HMM. Due to the commutativity of the summation and integration operations, the Baum functions for mixed-mode models all take the form of J-component mixtures, so that a mixed-mode model acts like a "bank" of single-mode CS-HMMs.

2.5.1 *Baum Probability Densities and Likelihood Evaluation*

The forward density is given by

$$\begin{aligned}\alpha(\mathbf{x}_n) &= p(\mathbf{Z}_n, \mathbf{x}_n) \\ &= \sum_{j=1}^{J} p(\mathbf{Z}_n, \mathbf{x}_n, j) \\ &= \sum_{j=1}^{J} p(\mathbf{Z}_n, \mathbf{x}_n | j)\, p(j) \\ &= \sum_{j=1}^{J} \rho_j\, \alpha_j(\mathbf{x}_n), \end{aligned} \tag{29}$$

where

$$\alpha_j(\mathbf{x}_n) = p(\mathbf{Z}_n, \mathbf{x}_n | j) \tag{30}$$

is the forward density obtained using the Baum recursion with the single-mode prior $p(\mathbf{x}_0|j)$. The backward density $\beta(\mathbf{x}_n)$ is the same for all j and is identical to the single-mode case. Integration of the terminal forward density over state space gives the measurement likelihood as a mixture of single-mode measurement likelihoods. Defining

$$p(\mathbf{Z}_N|j) = \int d\mathbf{x}_N\, \alpha_j(\mathbf{x}_N) \tag{31}$$

results in the likelihood being written as

$$p(\mathbf{Z}_N) = \sum_{j=1}^{J} \rho_j\, p(\mathbf{Z}_N|j). \tag{32}$$

The conditional mode-assignment probability

$$\rho_{j|N} = p(j|\mathbf{Z}_N) = \frac{1}{p(\mathbf{Z}_N)}\, p(\mathbf{Z}_N|j)\, \rho_j \tag{33}$$

is required along with the conditional state densities $\gamma(\mathbf{x}_n)$ and $\gamma(\mathbf{x}_n, \mathbf{x}_{n-1})$ to fully characterize the expected state evolution. Given $\rho_{j|N}$, the conditional state densities are

$$\begin{aligned}\gamma(\mathbf{x}_n) &= p(\mathbf{x}_n|\mathbf{Z}_N) \\ &= \sum_{j=1}^{J} p(\mathbf{x}_n, j|\mathbf{Z}_N) \\ &= \sum_{j=1}^{J} p(\mathbf{x}_n|j, \mathbf{Z}_N)\, p(j|\mathbf{Z}_N) \\ &= \sum_{j=1}^{J} \rho_{j|N}\, \gamma_j(\mathbf{x}_n), \end{aligned} \tag{34}$$

where

$$\begin{aligned}\gamma_j(\mathbf{x}_n) &= p(\mathbf{x}_n|\mathbf{Z}_N, j) \\ &= \frac{1}{p(\mathbf{Z}_N|j)}\,\alpha_j(\mathbf{x}_n)\,\beta(\mathbf{x}_n)\end{aligned} \tag{35}$$

is the conditional state density for a single-mode model with prior $p(\mathbf{x}_0|j)$. While maximization of equation (34) to obtain the optimal state sequence is a difficult non-linear optimization problem even for simple model densities, likelihood evaluation and parameter estimation (the two crucial problems for classification applications) can be performed directly from the densities. Finally, the joint state densities are

$$\begin{aligned}\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) &= p(\mathbf{x}_n, \mathbf{x}_{n-1}|\mathbf{Z}_N) \\ &= \sum_{j=1}^{J} p(\mathbf{x}_n, \mathbf{x}_{n-1}, j|\mathbf{Z}_N) \\ &= \sum_{j=1}^{J} p(\mathbf{x}_n, \mathbf{x}_{n-1}|j, \mathbf{Z}_N)\,p(j|\mathbf{Z}_N) \\ &= \sum_{j=1}^{J} \rho_{j|N}\,\gamma_j(\mathbf{x}_n, \mathbf{x}_{n-1}),\end{aligned} \tag{36}$$

where

$$\begin{aligned}\gamma_j(\mathbf{x}_n, \mathbf{x}_{n-1}) &= p(\mathbf{x}_n, \mathbf{x}_{n-1}|j, \mathbf{Z}_N) \\ &= \frac{1}{p(\mathbf{Z}_N|j)}\,\psi(\mathbf{x}_n)\,\delta_j(\mathbf{x}_n, \mathbf{x}_{n-1}).\end{aligned} \tag{37}$$

The expressions that are unique to mixed-mode models are summarized in table 6.

2.5.2 Parameter Estimation

For an observed measurement sequence, knowledge of the mode assignment variable j would reduce the mixed-mode modeling problem to a single-mode problem. The natural choice for the missing data in mixed-mode models therefore includes the mode assignment in addition to the state sequence. The resulting CDLF is

$$\begin{aligned}p(\mathcal{Z}, \mathcal{X}, \mathcal{J}) &= \prod_{k=1}^{K} p(\mathbf{Z}_{N_k}^k, \mathbf{X}_{N_k}^k, j_k) \\ &= \prod_{k=1}^{K} \rho_{j_k}\,p(\mathbf{x}_0^k|j_k) \prod_{n=1}^{N_k} p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k)\,p(\mathbf{z}_n^k|\mathbf{x}_n^k),\end{aligned}$$

where j_k is the mode assignment for the kth measurement and $\mathcal{J}$ represents the collection of mode assignments for all measurements.

Table 6. Baum Algorithm for Mixed-Mode CS-HMMs

Forward Density (Stage 1 Recursions)	$\delta_j(\mathbf{x}_n, \mathbf{x}_{n-1}) = p(\mathbf{x}_n \vert \mathbf{x}_{n-1})\, \alpha_j(\mathbf{x}_{n-1})$
Forward Density (Stage 2 Recursion)	$\alpha_j(\mathbf{x}_n) = p(\mathbf{z}_n \vert \mathbf{x}_n) \int d\mathbf{x}_{n-1}\, \delta_j(\mathbf{x}_n, \mathbf{x}_{n-1})$
Single-Mode Measurement Likelihood	$p(\mathbf{Z}_N \vert j) = \int d\mathbf{x}_N\, \alpha_j(\mathbf{x}_N)$
Measurement Likelihood	$p(\mathbf{Z}_N) = \sum_{j=1}^{J} \rho_j\, p(\mathbf{Z}_N \vert j)$
Conditional Mode-Assignment Probability	$\rho_{j\vert N} = \frac{1}{p(\mathbf{Z}_N)}\, \rho_j\, p(\mathbf{Z}_N \vert j)$
Conditional State Density	$\gamma(\mathbf{x}_n) = \frac{1}{p(\mathbf{Z}_N)} \sum_{j=1}^{J} \rho_{j\vert N}\, \alpha_j(\mathbf{x}_n)\, \beta(\mathbf{x}_n)$
Conditional Joint State Density	$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) = \frac{1}{p(\mathbf{Z}_N)} \sum_{j=1}^{J} \rho_{j\vert N}\, \delta_j(\mathbf{x}_n, \mathbf{x}_{n-1})\, \psi(\mathbf{x}_n)$

The EM auxiliary function in the mixed-mode case is

$$\begin{aligned} Q(\Theta, \Theta') &= E_{\mathcal{X},\mathcal{J}\vert\mathcal{Z},\Theta'}\Big\{ \log p(\mathcal{Z}, \mathcal{X}, \mathcal{J}) \Big\} \\ &= \sum_{\mathcal{J}} \int d\mathcal{X}\; p(\mathcal{X}, \mathcal{J} \vert \mathcal{Z}, \Theta')\; \log p(\mathcal{Z}, \mathcal{X}, \mathcal{J}) \\ &= \prod_{\ell=1}^{K} \sum_{j_\ell}^{J} \int d\mathbf{X}_{N_\ell}^{\ell}\; p(\mathbf{X}_{N_\ell}^{\ell}, j_\ell \vert \mathbf{Z}_{N_\ell}^{\ell}, \Theta')\; \log p(\mathcal{Z}, \mathcal{X}, \mathcal{J}). \end{aligned} \tag{38}$$

As before, the auxiliary function can be decomposed into components that depend exclusively on the different subsets of model parameters; that is,

$$Q(\Theta, \Theta') = Q_J(\rho, \Theta') + Q_0(\theta_0, \Theta') + Q_X(\theta_X, \Theta') + Q_Z(\theta_Z, \Theta'). \tag{39}$$

Components Q_X and Q_Z are identical in form to the single-mode counterparts because the relevant components from the CDLF (i.e., the product of transition densities for Q_X and of output densities for Q_Z) are independent of the mode assignment. Summation over j_ℓ thus serves to marginalize the mode assignments from the conditional density $p(\mathbf{X}_{N_\ell}^{\ell}, j_\ell \vert \mathbf{Z}_{N_\ell}^{\ell}, \Theta')$. The mixed-mode nature of the model shows up in the parameter estimates via the conditional state densities.

The auxiliary-function component Q_0 for the prior state density is

$$\begin{aligned} Q_0(\theta_0, \Theta^i) &= E_{\mathcal{X},\mathcal{J}|\mathcal{Z},\Theta^i}\left\{ \log \left[\prod_{k=1}^{K} p(\mathbf{x}_0^k | j_k) \right] \right\} \\ &= \sum_{j=1}^{J} \sum_{k=1}^{K} \int d\mathbf{x}_0^k \, \gamma(\mathbf{x}_0^k) \log p(\mathbf{x}_0^k | j). \end{aligned} \tag{40}$$

Component Q_J, which is used to optimize the mixing parameters, is given by

$$\begin{aligned} Q_J(\rho, \Theta^i) &= E_{\mathcal{X},\mathcal{J}|\mathcal{Z},\Theta^i}\left\{ \log \left[\prod_{k=1}^{K} \rho_{j_k} \right] \right\} \\ &= \prod_{\ell=1}^{K} \sum_{j_\ell=1}^{J} \int d\mathbf{X}_{N_\ell}^{\ell} \, p(\mathbf{X}_{N_\ell}^{\ell}, j_\ell | \mathbf{Z}_{N_\ell}^{\ell}, \Theta^i) \left\{ \log \left[\prod_{k=1}^{K} \rho_{j_k} \right] \right\} \\ &= \sum_{k=1}^{K} \sum_{j_k=1}^{J} \int d\mathbf{X}_{N_k}^{k} \, p(\mathbf{X}_{N_k}^{k}, j_k | \mathbf{Z}_{N_k}^{k}, \Theta^i) \log \rho_{j_k} \\ &= \sum_{k=1}^{K} \sum_{j_k=1}^{J} p(j_k | \mathbf{Z}_{N_k}^{k}, \Theta^i) \log \rho_{j_k} \\ &= \sum_{j=1}^{J} \sum_{k=1}^{K} \rho_{j_k|N_k} \log \rho_{j_k}. \end{aligned} \tag{41}$$

Since the model is linear in the mixing parameters, the EM update for these parameters can be expressed without knowing the form of the other model densities. The updates are obtained by maximizing Q_J, subject to the constraint that the ρ_j sum to one. The Lagrangian is

$$\tilde{Q}_J = \sum_{\ell=1}^{J} \sum_{k=1}^{K} \rho_{\ell_k|N_k} \log \rho_{\ell_k} + \lambda \left(1 - \sum_{\ell=1}^{J} \rho_\ell \right), \tag{42}$$

where ℓ is a dummy version of the mode assignment. Differentiating with respect to ρ_j and equating the resulting derivative to zero yields

$$\rho_j^{i+1} = \frac{1}{\lambda} \sum_{k=1}^{K} \rho_{j_k|N_k}. \tag{43}$$

Imposing the constraint $\sum_{j=1}^{J} \rho_j = 1$ gives the Lagrange multiplier as $\lambda = K$. The parameter update is therefore

$$\rho_j^{i+1} = \frac{1}{K} \sum_{k=1}^{K} \rho_{j_k|N_k}. \tag{44}$$

3. HIDDEN GAUSS-MARKOV MODELS

Hidden Gauss-Markov models (HGMMs) result when the model densities in the CS-HMM take the form

$$p(\mathbf{x}_n|\mathbf{x}_{n-1},\theta_X) = \mathcal{N}(\mathbf{x}_n;\mathbf{A}_n\mathbf{x}_{n-1},\mathbf{Q}_n) \quad (45)$$

$$p(\mathbf{z}_n|\mathbf{x}_n,\theta_Z) = \mathcal{N}(\mathbf{z}_n;\mathbf{B}_n\mathbf{x}_n,\mathbf{R}_n) \quad (46)$$

$$p(\mathbf{x}_0|\theta_0) = \mathcal{N}(\mathbf{x}_0;\mu_0,\mathbf{P}_0)\,, \quad (47)$$

where the usual shorthand $\mathcal{N}(\mathbf{y};\mu,\mathbf{P})$ is used to denote the density function for a multivariate normal vector $\mathbf{y}$ with mean μ and covariance matrix $\mathbf{P}$. For example, if $\mathbf{y}$ is L-dimensional, then the density function is

$$\mathcal{N}(\mathbf{y};\mu,\mathbf{P}) = (2\pi)^{-L/2}\,|\mathbf{P}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mu)^{\mathrm{T}}\,\mathbf{P}^{-1}\,(\mathbf{y}-\mu)\right\}. \quad (48)$$

The model densities in an HGMM are parameterized by the sets $\theta_0 = \{\mu_0,\mathbf{P}_0\}$, $\theta_X = \{\mathbf{A}_n,\mathbf{Q}_n\}$, and $\theta_Z = \{\mathbf{B}_n,\mathbf{R}_n\}$, where $n = 1,\ldots,N$. The transition matrices $\{\mathbf{A}_n\}$, output matrices $\{\mathbf{B}_n\}$, and covariance matrices $\{\mathbf{Q}_n\}$ and $\{\mathbf{R}_n\}$ are collectively referred to as the *system matrices*. Although time-varying models are considered throughout most of this section for the sake of generality, time-invariant models (whose system matrices are the same for all n) are discussed briefly in section 3.5, where a parameter-invariance structure is noted for the measurement likelihood function.

The model densities that characterize HGMMs provide alternative expressions for the set of equations defined by

$$\mathbf{x}_n = \mathbf{A}_n\,\mathbf{x}_{n-1} + \mathbf{w}_n \quad (49)$$

$$\mathbf{z}_n = \mathbf{B}_n\,\mathbf{x}_n + \mathbf{v}_n \quad (50)$$

$$p(\mathbf{w}_n) = \mathcal{N}(\mathbf{w}_n;\mathbf{0},\mathbf{Q}_n) \quad (51)$$

$$p(\mathbf{v}_n) = \mathcal{N}(\mathbf{v}_n;\mathbf{0},\mathbf{R}_n) \quad (52)$$

$$p(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0;\mu_0,\mathbf{P}_0)\,, \quad (53)$$

which is recognized as the defining model for the Kalman filter. While Kalman filters are not typically viewed in the context of HMMs, they have recently been described as being "analogous" to CS-HMMs [25, 27, 35]. This section demonstrates that the relationship is not merely an analogy, but that Kalman-filter models in fact form a subset of CS-HMMs. The CS-HMM results given in the previous section are specialized to the Gaussian models in equations (45), (46), and (47) to show that Baum's forward-backward algorithm and the Viterbi algorithm are implemented by the two-filter [42, 43]

and RTS [41] formulations of the fixed-interval Kalman smoother, respectively. The measurement likelihood obtained from the forward pass of the Baum algorithm is shown to equal the innovation-based definition from Kalman-filter theory [44], and an existing EM parameter estimation algorithm [31, 32] is shown to follow directly from the CS-HMM auxiliary function.

Also, paralleling the developments of the previous section, mixed-mode HGMMs are defined in which the single-component prior density in equation (47) is replaced by the J-component mixture

$$p(\mathbf{x}_0|\theta_0,\rho) \;=\; \sum_{j=1}^{J} \rho_j \, \mathcal{N}\left(\mathbf{x}_0; \mu_0^j, \mathbf{P}_0^j\right), \tag{54}$$

where the mixing parameters satisfy $\rho_j \geq 0$ for all j and $\sum_{j=1}^{J} \rho_j = 1$. Here, $\theta_0 = \left\{\mu_0^j, \mathbf{P}_0^j, j = 1, \ldots, J\right\}$ contains the parameters for each mode in the mixture. The mixed-mode HGMM is developed to provide more flexible and accurate models for short measurement sequences whose assessed likelihoods are very sensitive to the prior distribution, and to better represent classes of signals whose members are well modeled by the same set of system matrices, but which exhibit significant within-class variability due to different initial-state values.

3.1 GAUSSIAN REFACTORIZATION LEMMA

Derivation of the HGMM algorithms is simplified by introducing the following Gaussian refactorization lemma (GRL), whose proof is given in appendix A.

Lemma: Given the L-dimensional vector $\mathbf{x}$, the M-dimensional vector $\mathbf{y}$, appropriately sized nonsingular covariance matrices $\mathbf{S}$ and $\mathbf{P}$, and the $M \times L$ matrix $\mathbf{F}$, the product function

$$\eta(\mathbf{y},\mathbf{x}) \;=\; \mathcal{N}(\mathbf{y}; \mathbf{F}\mathbf{x}, \mathbf{S})\, \mathcal{N}(\mathbf{x}; \mu, \mathbf{P}) \tag{55}$$

can be refactored as

$$\eta(\mathbf{y},\mathbf{x}) \;=\; \mathcal{N}(\mathbf{y}; \omega, \mathbf{\Omega})\, \mathcal{N}(\mathbf{x}; \lambda, \mathbf{\Lambda}), \tag{56}$$

where the means and covariances in the resulting product densities are

$$\omega \;=\; \mathbf{F}\mu \tag{57}$$

$$\mathbf{\Omega} \;=\; \mathbf{S} + \mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}} \tag{58}$$

$$\lambda \;=\; (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mu + \mathbf{H}\mathbf{y} \tag{59}$$

$$\mathbf{\Lambda} \;=\; (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mathbf{P}, \tag{60}$$

with the supporting variable

$$\mathbf{H} = \mathbf{P}\mathbf{F}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}. \tag{61}$$

The parameters can also be expressed in information terms, where the *information matrix* is the inverse of the covariance matrix and the *information vector* is the mean vector premultiplied by the information matrix. In this format, the density parameters are

$$\boldsymbol{\Lambda}^{-1}\lambda = \mathbf{P}^{-1}\mu + \mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} \tag{62}$$

$$\boldsymbol{\Lambda}^{-1} = \mathbf{P}^{-1} + \mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{F} \tag{63}$$

$$\boldsymbol{\Omega}^{-1}\omega = \mathbf{D}\mathbf{P}^{-1}\mu \tag{64}$$

$$\boldsymbol{\Omega}^{-1} = \left(\mathbf{I} - \mathbf{D}\mathbf{F}^{\mathrm{T}}\right)\mathbf{S}^{-1}, \tag{65}$$

with the supporting variable

$$\mathbf{D} = \mathbf{S}^{-1}\mathbf{F}\left(\boldsymbol{\Lambda}^{-1}\right)^{-1}. \tag{66}$$

Interestingly, while this lemma arises as a necessary prerequisite for evaluating the density recursions in the Baum algorithm, it naturally generates all of the Kalman-filter update recursions.

3.2 BAUM ALGORITHM

This subsection applies the results summarized in table 4 to models with the Gaussian densities in equations (45), (46), and (47). The derivations for the forward and backward recursions proceed as an induction. The conditional state densities and likelihood are then obtained in terms of the forward and backward densities.

3.2.1 Forward Densities

The assumed form for the forward density at time t_{n-1} is

$$\alpha(\mathbf{x}_{n-1}) = c_{n-1}\,\mathcal{N}\left(\mathbf{x}_{n-1}; \mu_{n-1|n-1}, \mathbf{P}_{n-1|n-1}\right),$$

which includes the initial condition by letting $\mathbf{P}_{0|0} = \mathbf{P}_0$, $\mu_{0|0} = \mu_0$, and $c_0 = 1$. The first stage of the forward recursion evaluates the term

$$\begin{aligned}\delta(\mathbf{x}_n, \mathbf{x}_{n-1}) &= p(\mathbf{x}_n|\mathbf{x}_{n-1})\,\alpha(\mathbf{x}_{n-1}) \\ &= c_{n-1}\,\mathcal{N}(\mathbf{x}_n; \mathbf{A}_n\mathbf{x}_{n-1}, \mathbf{Q}_n)\,\mathcal{N}\left(\mathbf{x}_{n-1}; \mu_{n-1|n-1}, \mathbf{P}_{n-1|n-1}\right).\end{aligned}$$

Application of the GRL results in

$$\delta(\mathbf{x}_n, \mathbf{x}_{n-1}) = c_{n-1}\,\mathcal{N}\left(\mathbf{x}_n; \mu_{n|n-1}, \mathbf{P}_{n|n-1}\right)\,\mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \Lambda_n)\,, \tag{67}$$

where the mean and covariance parameters in the first factor correspond to a Kalman-filter time update and are given by

$$\mu_{n|n-1} = \mathbf{A}_n \mu_{n-1|n-1} \tag{68}$$

$$\mathbf{P}_{n|n-1} = \mathbf{Q}_n + \mathbf{A}_n \mathbf{P}_{n-1|n-1} \mathbf{A}_n^{\mathrm{T}}, \tag{69}$$

respectively. The parameters in the second factor are

$$\lambda_n = (\mathbf{I} - \mathbf{H}_n \mathbf{A}_n)\,\mu_{n-1|n-1} + \mathbf{H}_n \mathbf{x}_n \tag{70}$$

$$\Lambda_n = (\mathbf{I} - \mathbf{H}_n \mathbf{A}_n)\,\mathbf{P}_{n-1|n-1}, \tag{71}$$

where

$$\mathbf{H}_n = \mathbf{P}_{n-1|n-1} \mathbf{A}_n^{\mathrm{T}} \mathbf{P}_{n|n-1}^{-1}. \tag{72}$$

These variables are usually ignored in a Kalman-filter context, but, in an HGMM context, they occur again in the smoothing and parameter-estimation problems. While the mean λ_n of the second term is a function of the current state $\mathbf{x}_n$, integration of this term over $\mathbf{x}_{n-1}$ produces unity regardless of the mean; that is,

$$\int d\mathbf{x}_{n-1}\, \mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \Lambda_n) = 1.$$

The forward density then becomes

$$\begin{aligned} \alpha(\mathbf{x}_n) &= p(\mathbf{z}_n|\mathbf{x}_n) \int d\mathbf{x}_{n-1}\, \delta(\mathbf{x}_n, \mathbf{x}_{n-1}) \\ &= c_{n-1}\,\mathcal{N}(\mathbf{z}_n; \mathbf{B}_n \mathbf{x}_n, \mathbf{R}_n)\,\mathcal{N}\left(\mathbf{x}_n; \mu_{n|n-1}, \mathbf{P}_{n|n-1}\right). \end{aligned}$$

Applying the GRL to this product yields

$$\alpha(\mathbf{x}_n) = c_{n-1}\,\mathcal{N}(\mathbf{z}_n; \hat{\mathbf{z}}_n, \boldsymbol{\Sigma}_n)\,\mathcal{N}\left(\mathbf{x}_n; \mu_{n|n}, \mathbf{P}_{n|n}\right), \tag{73}$$

where the parameters in the first factor are the estimated measurement and its error covariance, respectively given by

$$\hat{\mathbf{z}}_n = \mathbf{B}_n \mu_{n|n-1} \tag{74}$$

$$\boldsymbol{\Sigma}_n = \mathbf{R}_n + \mathbf{B}_n \mathbf{P}_{n|n-1} \mathbf{B}_n^{\mathrm{T}}. \tag{75}$$

The parameters in the second factor correspond to the Kalman-filter measurement update and are

$$\mu_{n|n} = (\mathbf{I} - \mathbf{G}_n \mathbf{B}_n)\, \mu_{n|n-1} + \mathbf{G}_n \mathbf{z}_n \tag{76}$$

$$\mathbf{P}_{n|n} = (\mathbf{I} - \mathbf{G}_n \mathbf{B}_n)\, \mathbf{P}_{n|n-1}, \tag{77}$$

where

$$\mathbf{G}_n = \mathbf{P}_{n|n-1} \mathbf{B}_n^{\mathrm{T}} \boldsymbol{\Sigma}_n^{-1}. \tag{78}$$

Defining the innovation vector as

$$\nu_n = \mathbf{z}_n - \hat{\mathbf{z}}_n = \mathbf{z}_n - \mathbf{B}_n\, \mu_{n|n-1} \tag{79}$$

and recursively defining the weighting constant as

$$c_n = c_{n-1}\, \mathcal{N}(\nu_n; \mathbf{0}, \boldsymbol{\Sigma}_n) \tag{80}$$

results in equation (73) being rewritten as

$$\alpha(\mathbf{x}_n) = c_n\, \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n}, \mathbf{P}_{n|n}\right), \tag{81}$$

which matches the assumed form for the previous time step, completing the induction.

3.2.2 Likelihood Evaluation

The likelihood of a measurement sequence is obtained using equation (13), that is, by integrating the forward density at time t_N over all of state space, giving

$$\begin{aligned} p(\mathbf{Z}_N) &= \int d\mathbf{x}_N\, \alpha(\mathbf{x}_N) = c_N \\ &= \prod_{n=1}^{N} \mathcal{N}(\nu_n; \mathbf{0}, \boldsymbol{\Sigma}_n), \end{aligned} \tag{82}$$

which equals the known likelihood expression for the Kalman filter [44]. This definition for the likelihood can also be applied to the partial sequence $\mathbf{Z}_n$ to obtain $p(\mathbf{Z}_n) = c_n$. Since the forward density can be decomposed as

$$\begin{aligned} \alpha(\mathbf{x}_n) &= p(\mathbf{Z}_n, \mathbf{x}_n) \\ &= p(\mathbf{Z}_n)\, p(\mathbf{x}_n | \mathbf{Z}_n), \end{aligned} \tag{83}$$

it follows that

$$p(\mathbf{x}_n | \mathbf{Z}_n) = \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n}, \mathbf{P}_{n|n}\right). \tag{84}$$

The "joint forward density" $\delta(\mathbf{x}_n, \mathbf{x}_{n-1})$ similarly decomposes as

$$\begin{aligned}\delta(\mathbf{x}_n, \mathbf{x}_{n-1}) &= p(\mathbf{Z}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1}) \\ &= p(\mathbf{Z}_{n-1})\, p(\mathbf{x}_n|\mathbf{Z}_{n-1})\, p(\mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{Z}_{n-1}),\end{aligned} \tag{85}$$

where

$$p(\mathbf{x}_n|\mathbf{Z}_{n-1}) = \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n-1}, \mathbf{P}_{n|n-1}\right) \tag{86}$$

$$p(\mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{Z}_{n-1}) = \mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \Lambda_n). \tag{87}$$

3.2.3 Backward Densities

The assumed form for backward density is

$$\beta(\mathbf{x}_n) = c_n^{(\mathrm{r})} \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n+1}^{(\mathrm{r})}, \mathbf{P}_{n|n+1}^{(\mathrm{r})}\right),$$

where the superscript (r) indicates reverse time. The recursion again proceeds in two stages, with the first stage evaluating the product

$$\begin{aligned}\psi(\mathbf{x}_n) &= p(\mathbf{z}_n|\mathbf{x}_n)\, \beta(\mathbf{x}_n) \\ &= c_n^{(\mathrm{r})} \mathcal{N}(\mathbf{z}_n; \mathbf{B}_n\mathbf{x}_n, \mathbf{R}_n)\, \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n+1}^{(\mathrm{r})}, \mathbf{P}_{n|n+1}^{(\mathrm{r})}\right).\end{aligned}$$

Application of the GRL produces

$$\psi(\mathbf{x}_n) = c_n^{(\mathrm{r})} \mathcal{N}\left(\mathbf{z}_n; \hat{\mathbf{z}}_n^{(\mathrm{r})}, \Sigma_n^{(\mathrm{r})}\right) \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n}^{(\mathrm{r})}, \mathbf{P}_{n|n}^{(\mathrm{r})}\right), \tag{88}$$

where the mean and covariance of the first factor are the reverse-time measurement estimate and its error covariance, given by

$$\hat{\mathbf{z}}_n^{(\mathrm{r})} = \mathbf{B}_n \mu_{n|n+1}^{(\mathrm{r})} \tag{89}$$

$$\Sigma_n^{(\mathrm{r})} = \mathbf{R}_n + \mathbf{B}_n \mathbf{P}_{n|n+1}^{(\mathrm{r})} \mathbf{B}_n^{\mathrm{T}}. \tag{90}$$

respectively. The parameters in the second factor, which correspond to a reverse-time Kalman-filter measurement update, are

$$\mu_{n|n}^{(\mathrm{r})} = \left(\mathbf{I} - \mathbf{G}_n^{(\mathrm{r})}\mathbf{B}_n\right) \mu_{n|n+1}^{(\mathrm{r})} + \mathbf{G}_n^{(\mathrm{r})}\mathbf{z}_n \tag{91}$$

$$\mathbf{P}_{n|n}^{(\mathrm{r})} = \left(\mathbf{I} - \mathbf{G}_n^{(\mathrm{r})}\mathbf{B}_n\right) \mathbf{P}_{n|n+1}^{(\mathrm{r})}, \tag{92}$$

where

$$\mathbf{G}_n^{(\mathrm{r})} = \mathbf{P}_{n|n+1}^{(\mathrm{r})} \mathbf{B}_n^{\mathrm{T}} \Sigma_n^{(\mathrm{r})^{-1}}. \tag{93}$$

Defining the weighting-constant update

$$c_{n-1}^{(r)} = c_n^{(r)} \left| \mathbf{A}_n \right|^{-1} \mathcal{N}\left(\mathbf{z}_n; \hat{\mathbf{z}}_n^{(r)}, \boldsymbol{\Sigma}_n^{(r)}\right) \tag{94}$$

yields

$$\psi(\mathbf{x}_n) = c_{n-1}^{(r)} \left| \mathbf{A}_n \right| \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n}^{(r)}, \mathbf{P}_{n|n}^{(r)}\right). \tag{95}$$

Here, the canceling terms $|\mathbf{A}_n^{-1}|$ and $|\mathbf{A}_n|$ have been inserted to accommodate a necessary factor in the second stage of the recursion, which evaluates the integral

$$\begin{aligned} \beta(\mathbf{x}_{n-1}) &= \int d\mathbf{x}_n \; p(\mathbf{x}_n|\mathbf{x}_{n-1})\, \psi(\mathbf{x}_n) \\ &= c_{n-1}^{(r)} \int d\mathbf{x}_n \, \eta(\mathbf{x}_n, \mathbf{x}_{n-1}), \end{aligned} \tag{96}$$

where

$$\eta(\mathbf{x}_n, \mathbf{x}_{n-1}) = \left| \mathbf{A}_n \right| \mathcal{N}(\mathbf{x}_n; \mathbf{A}_n \mathbf{x}_{n-1}, \mathbf{Q}_n)\, \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n}^{(r)}, \mathbf{P}_{n|n}^{(r)}\right). \tag{97}$$

This product does not immediately fit the form required for application of the GRL. If $\mathbf{A}_n$ is invertible, however, then

$$\mathcal{N}\left(\mathbf{x}_{n-1}; \mathbf{A}_n^{-1}\mathbf{x}_n, \mathbf{A}_n^{-1}\mathbf{Q}_n\mathbf{A}_n^{-\mathrm{T}}\right) = \left| \mathbf{A}_n \right| \mathcal{N}(\mathbf{x}_n; \mathbf{A}_n \mathbf{x}_{n-1}, \mathbf{Q}_n). \tag{98}$$

The GRL can then be used to obtain

$$\begin{aligned} \eta(\mathbf{x}_n, \mathbf{x}_{n-1}) &= \mathcal{N}\left(\mathbf{x}_{n-1}; \mathbf{A}_n^{-1}\mathbf{x}_n, \mathbf{A}_n^{-1}\mathbf{Q}_n\mathbf{A}_n^{-\mathrm{T}}\right) \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n}^{(r)}, \mathbf{P}_{n|n}^{(r)}\right) \\ &= \mathcal{N}\left(\mathbf{x}_{n-1}; \mu_{n-1|n}^{(r)}, \mathbf{P}_{n-1|n}^{(r)}\right) \mathcal{N}\left(\mathbf{x}_n; \lambda_n^{(r)}, \boldsymbol{\Lambda}_n^{(r)}\right). \end{aligned} \tag{99}$$

The parameters in the first factor correspond to a reverse-time Kalman-filter time update, given by

$$\mu_{n-1|n}^{(r)} = \mathbf{A}_n^{-1} \mu_{n|n}^{(r)} \tag{100}$$

$$\mathbf{P}_{n-1|n}^{(r)} = \mathbf{A}_n^{-1} \left(\mathbf{Q}_n + \mathbf{P}_{n|n}^{(r)}\right) \mathbf{A}_n^{-\mathrm{T}}. \tag{101}$$

The parameters in the second factor are

$$\lambda_n^{(r)} = \left(\mathbf{I} - \mathbf{H}_n^{(r)} \mathbf{A}_n^{-1}\right) \mu_{n|n}^{(r)} + \mathbf{H}_n^{(r)} \mathbf{x}_{n-1} \tag{102}$$

$$\boldsymbol{\Lambda}_n^{(r)} = \left(\mathbf{I} - \mathbf{H}_n^{(r)} \mathbf{A}_n^{-1}\right) \mathbf{P}_{n|n}^{(r)}, \tag{103}$$

where

$$\mathbf{H}_n^{(r)} = \mathbf{P}_{n|n}^{(r)} \left(\mathbf{Q}_n + \mathbf{P}_{n|n}^{(r)}\right)^{-1} \mathbf{A}_n. \tag{104}$$

Substituting the refactored product into equation (96) and integrating results in

$$\beta(\mathbf{x}_{n-1}) \;=\; c_{n-1}^{(r)}\,\mathcal{N}\!\left(\mathbf{x}_{n-1};\mu_{n-1|n}^{(r)},\mathbf{P}_{n-1|n}^{(r)}\right). \tag{105}$$

The induction requires the assumed form to fit the initial condition, but this is not the case since $\beta(\mathbf{x}_N) = 1$ is not a Gaussian density. As noted in subsection 2.1.2, the recursion should really be started at time t_{N-1}, where it is evaluated as a special case using

$$\beta(\mathbf{x}_{N-1}) \;=\; \int d\mathbf{x}_N\,\mathcal{N}(\mathbf{z}_N;\mathbf{B}_N\mathbf{x}_N,\mathbf{R}_N)\,\mathcal{N}(\mathbf{x}_N;\mathbf{A}_N\mathbf{x}_{N-1},\mathbf{Q}_N). \tag{106}$$

To avoid having to consider this special case in all of the recursions dependent on $\beta(\mathbf{x}_n)$, an approach is taken here that parallels the development of the two-filter Kalman smoother [42, 43]. That is, an information formulation for the reverse-time filter is used, and the terminal-state parameters are defined as $\mathbf{P}_{N|N+1}^{(r)-1} = \mathbf{0}$ and $\mathbf{P}_{N|N+1}^{(r)-1}\,\mu_{N|N+1} = \mathbf{0}$. This approach is equivalent to defining $\beta(\mathbf{x}_N)$ as a Gaussian "pseudo-density" whose variances are infinite but whose value for any argument is unity. The information formulation of the reverse-time filter is reflected in table 7, which summarizes the computations involved in the Baum algorithm.

3.2.4 Conditional State Densities: Method I

The state density $\gamma(\mathbf{x}_n)$ is the normalized product of $\alpha(\mathbf{x}_n)$ and $\beta(\mathbf{x}_n)$, which, by definition, gives a properly normalized density. When the Gaussian densities in $\alpha(\mathbf{x}_n)$ and $\beta(\mathbf{x}_n)$ are multiplied without the scale constants, the product is a properly normalized density function. The scale constants c_n and $c_n^{(r)}$ can therefore be ignored when constructing the $\gamma(\mathbf{x}_n)$, giving

$$\gamma(\mathbf{x}_n) \;=\; \mathcal{N}\!\left(\mathbf{x}_n;\mu_{n|n},\mathbf{P}_{n|n}\right)\mathcal{N}\!\left(\mathbf{x}_n;\mu_{n|n+1}^{(r)},\mathbf{P}_{n|n+1}^{(r)}\right).$$

This is a product of Gaussians in the same variable with constant means and covariances, which has the following well-known form:

$$\gamma(\mathbf{x}_n) \;=\; \mathcal{N}\!\left(\mathbf{x}_n;\mu_{n|N},\mathbf{P}_{n|N}\right), \tag{107}$$

where

$$\mathbf{P}_{n|N} \;=\; \left(\mathbf{P}_{n|n}^{-1} + \mathbf{P}_{n|n+1}^{(r)\,-1}\right)^{-1} \tag{108}$$

$$\mu_{n|N} \;=\; \mathbf{P}_{n|N}\left(\mathbf{P}_{n|n}^{-1}\,\mu_{n|n} + \mathbf{P}_{n|n+1}^{(r)\,-1}\,\mu_{n|n+1}^{(r)}\right). \tag{109}$$

Table 7. Baum Algorithm for HGMMs

Forward Stage 1 Recursion (Time Update) Given: $\mu_{n-1\|n-1}, \mathbf{P}_{n-1\|n-1}$	$\mu_{n\|n-1} = \mathbf{A}_n \mu_{n-1\|n-1}$ $\mathbf{P}_{n\|n-1} = \mathbf{Q}_n + \mathbf{A}_n \mathbf{P}_{n-1\|n-1} \mathbf{A}_n^{\mathrm{T}}$ $\mathbf{H}_n = \mathbf{P}_{n-1\|n-1} \mathbf{A}_n^{\mathrm{T}} \mathbf{P}_{n\|n-1}^{-1}$
Forward Stage 2 Recursion (Measurement Update) Given: $\mu_{n\|n-1}, \mathbf{P}_{n\|n-1}$	$\nu_n = \mathbf{z}_n - \mathbf{B}_n \mu_{n\|n-1}$ $\mathbf{\Sigma}_n = \mathbf{R}_n + \mathbf{B}_n \mathbf{P}_{n\|n-1} \mathbf{B}_n^{\mathrm{T}}$ $\mathbf{G}_n = \mathbf{P}_{n\|n-1} \mathbf{B}_n^{\mathrm{T}} \mathbf{\Sigma}_n^{-1}$ $\mu_{n\|n} = (\mathbf{I} - \mathbf{G}_n \mathbf{B}_n)\, \mu_{n\|n-1} + \mathbf{G}_n \mathbf{z}_n$ $\mathbf{P}_{n\|n} = (\mathbf{I} - \mathbf{G}_n \mathbf{B}_n)\, \mathbf{P}_{n\|n-1}$
Measurement Likelihood	$p(\mathbf{Z}) = \prod_{n=1}^{N} \mathcal{N}(\nu_n; \mathbf{0}, \mathbf{\Sigma}_n)$
Backward Information Variables	$\xi_n = \mathbf{P}_n^{(r)-1} \mu_n^{(r)}, \;\; \mathbf{\Gamma}_n = \mathbf{P}_n^{(r)-1}$
Backward Stage 1 Recursion (Measurement Update) Given: $\xi_{n\|n+1}, \mathbf{\Gamma}_{n\|n+1}$	$\xi_{n\|n} = \xi_{n\|n+1} + \mathbf{B}_n^{\mathrm{T}} \mathbf{R}_n^{-1} \mathbf{z}_n$ $\mathbf{\Gamma}_{n\|n} = \mathbf{\Gamma}_{n\|n+1} + \mathbf{B}_n^{\mathrm{T}} \mathbf{R}_n^{-1} \mathbf{B}_n$
Backward Stage 2 Recursion (Time Update) Given: $\xi_{n\|n}, \mathbf{\Gamma}_{n\|n}$	$\xi_{n-1\|n} = \mathbf{A}_n^{\mathrm{T}} \mathbf{Q}_n^{-1} \left(\mathbf{\Gamma}_{n\|n} + \mathbf{Q}_n^{-1}\right)^{-1} \xi_{n\|n}$ $\mathbf{\Gamma}_{n-1\|n} = \mathbf{A}_n^{\mathrm{T}} \left[\mathbf{Q}_n^{-1} - \mathbf{Q}_n^{-1} \left(\mathbf{\Gamma}_{n\|n} + \mathbf{Q}_n^{-1}\right)^{-1} \mathbf{Q}_n^{-1}\right] \mathbf{A}_n$
Conditional State Calculation (Smoothing) Given: $\mu_{n\|n}, \mathbf{P}_{n\|n}, \xi_{n\|n+1}, \mathbf{\Gamma}_{n\|n+1}$	$\mathbf{P}_{n\|N} = \left(\mathbf{P}_{n\|n}^{-1} + \mathbf{\Gamma}_{n\|n+1}\right)^{-1}$ $\mu_{n\|N} = \mathbf{P}_{n\|N}^{-1} \left(\mathbf{P}_{n\|n}^{-1} \mu_{n\|n} + \xi_{n\|n+1}\right)$
Adjacent-State Cross Covariance Given: $\mathbf{P}_{n\|N}, \mathbf{H}_n$	$\mathbf{P}_{n,n-1\|N} = \mathbf{P}_{n\|N} \mathbf{H}_n^{\mathrm{T}}$

The use of the information format for the backward density parameters simplifies these calculations since the variables $\mathbf{P}_{n|n+1}^{(r)\,-1}$ and $\mathbf{P}_{n|n+1}^{(r)\,-1}\, \mu_{n|n+1}^{(r)}$ are generated by the backward recursions and need not be calculated from the covariance information.

As outlined above, the calculations involved in obtaining these densities via Baum's forward-backward algorithm are exactly the same calculations involved in the two-filter implementation of the fixed-interval Kalman smoother [42, 43]. The two-filter smoothing algorithm is therefore an implementation of the Baum algorithm for HGMMs.

3.2.5 *Conditional Joint State Densities*

Characterization of the conditional state evolution for HGMMs is completed by specifying the joint density of time-adjacent states, which is obtained from equation (10) by substituting equations (67) and (95). This calculation gives

$$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) = \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n-1}, \mathbf{P}_{n|n-1}\right) \mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \mathbf{\Lambda}_n) \mathcal{N}\left(\mathbf{x}_n; \mu^{(\mathrm{r})}_{n|n}, \mathbf{P}^{(\mathrm{r})}_{n|n}\right), \quad (110)$$

where the means and covariances of these factors are defined in equations (68), (69), (70), (71), (91), and (92). It is easy to show that the product of the first and third terms provides an alternative construction for $\gamma(\mathbf{x}_n)$, such that

$$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) = \mathcal{N}\left(\mathbf{x}_n; \mu_{n|N}, \mathbf{P}_{n|N}\right) \mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \mathbf{\Lambda}_n). \quad (111)$$

The accuracy of this expression is confirmed by recalling that $\mathcal{N}\left(\mathbf{x}_n; \mu_{n|N}, \mathbf{P}_{n|N}\right) = p(\mathbf{x}_n|\mathbf{Z}_N)$ and that $\mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \mathbf{\Lambda}_n) = p(\mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{Z}_{n-1})$. In the latter expression, the conditioning variable can be changed from $\mathbf{Z}_{n-1}$ to $\mathbf{Z}_N$ since the information contained in $\mathbf{Z}^{\mathrm{C}}_{n-1}$ is redundant, given that conditioning on $\mathbf{x}_n$ also occurs, so that

$$\mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \mathbf{\Lambda}_n) = p(\mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{Z}_N). \quad (112)$$

The product in equation (111) is therefore

$$\begin{aligned} \gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) &= p(\mathbf{x}_n|\mathbf{Z}_N)\, p(\mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{Z}_N) \\ &= p(\mathbf{x}_n, \mathbf{x}_{n-1}|\mathbf{Z}_N), \end{aligned} \quad (113)$$

which is the desired definition. Equation (111) is instrumental for evaluating the auxiliary function for HGMMs.

For the computations actually performed during parameter estimation, only the cross-covariance matrix for time-adjacent states, denoted $\mathbf{P}_{n,n-1|N}$, is required. An expression for this matrix can be obtained by evaluating the conditional joint density and then extracting the cross-covariance matrix from one of the off-diagonal blocks of the joint covariance matrix. This derivation is provided in appendix B, where it is shown that the density of the $2L \times 1$ joint random vector $\mathbf{x}_{[n,n-1]} = [\mathbf{x}^{\mathrm{T}}_n, \mathbf{x}^{\mathrm{T}}_{n-1}]^{\mathrm{T}}$ is given by

$$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) = \mathcal{N}\left(\mathbf{x}_{[n,n-1]}; \mu_{[n,n-1]|N}, \mathbf{P}_{[n,n-1]|N}\right), \quad (114)$$

where

$$\mu_{[n,n-1]|N} = \begin{bmatrix} \mu_{n|N} \\ \mu_{n-1|N} \end{bmatrix} \quad (115)$$

$$\mathbf{P}_{[n,n-1]|N} = \begin{bmatrix} \mathbf{P}_{n|N} & \mathbf{P}_{n|N}\mathbf{H}^{\mathrm{T}}_n \\ \mathbf{H}_n\mathbf{P}_{n|N} & \mathbf{P}_{n-1|N} \end{bmatrix}. \quad (116)$$

The term $\mathbf{H}_n$ is defined in equation (72). The upper off-diagonal gives the adjacent-state cross-covariance matrix as

$$\mathbf{P}_{n,n-1|N} = \mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}. \tag{117}$$

This expression is considerably simpler than the recursive definition given by Shumway and Stoffer [31].

3.2.6 Conditional State Densities: Method II

The expression for the conditional joint state density in equation (111) provides an alternative approach to calculating the conditional state densities. In particular, the conditional state density can be obtained by (1) calculating the forward densities $\alpha(\mathbf{x}_n)$ for $n = 1, \ldots, N$, (2) initializing $\gamma(\mathbf{x}_N) = \alpha(\mathbf{x}_N)$, and then (3) calculating $\gamma(\mathbf{x}_{n-1})$ given $\gamma(\mathbf{x}_n)$ for $n = N, \ldots, 1$. To realize this algorithm, the backward recursion defined in step (3) must be derived. This derivation begins by noting that

$$\begin{aligned} \gamma(\mathbf{x}_{n-1}) &= \int d\mathbf{x}_n \, \gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) \\ &= \int d\mathbf{x}_n \, \mathcal{N}\left(\mathbf{x}_n; \mu_{n|N}, \mathbf{P}_{n|N}\right) \mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \Lambda_n), \end{aligned} \tag{118}$$

where the expression for the joint density given in equation (111) has been substituted. For convenience, the definitions of λ_n and Λ_n are restated as

$$\begin{aligned} \lambda_n &= (\mathbf{I} - \mathbf{H}_n\mathbf{A}_n)\,\mu_{n-1|n-1} + \mathbf{H}_n\mathbf{x}_n \\ \Lambda_n &= (\mathbf{I} - \mathbf{H}_n\mathbf{A}_n)\,\mathbf{P}_{n-1|n-1}, \end{aligned}$$

where the intermediate variable $\mathbf{H}_n$ is

$$\mathbf{H}_n = \mathbf{P}_{n-1|n-1}\mathbf{A}_n^{\mathrm{T}}\mathbf{P}_{n|n-1}^{-1}.$$

Given these variable definitions, the mean in the second factor in equation (118) is seen to have an "offset" of $(\mathbf{I} - \mathbf{H}_n\mathbf{A}_n)\,\mu_{n-1|n-1}$. This offset is temporarily removed by defining the new variable

$$\mathbf{y}_{n-1} = \mathbf{x}_{n-1} - (\mathbf{I} - \mathbf{H}_n\mathbf{A}_n)\,\mu_{n-1|n-1}. \tag{119}$$

With this, the product in equation (118) can be rewritten as

$$\gamma(\mathbf{x}_{n-1}) = \int d\mathbf{x}_n \, \mathcal{N}\left(\mathbf{x}_n; \mu_{n|N}, \mathbf{P}_{n|N}\right) \mathcal{N}(\mathbf{y}_{n-1}; \mathbf{H}_n\mathbf{x}_n, \Lambda_n). \tag{120}$$

Application of the GRL then gives

$$\gamma(\mathbf{x}_{n-1}) = \int d\mathbf{x}_n \, \mathcal{N}(\mathbf{x}_n; \upsilon_n, \Upsilon_n) \, \mathcal{N}(\mathbf{y}_{n-1}; \omega_n, \Omega_n). \tag{121}$$

The mean and covariance of the new $\mathbf{y}_{n-1}$ term are

$$\omega_n = \mathbf{H}_n \mu_{n|N} \tag{122}$$

$$\Omega_n = \Lambda_n + \mathbf{H}_n \mathbf{P}_{n|N} \mathbf{H}_n^{\mathrm{T}}, \tag{123}$$

respectively. While the mean υ_n and covariance Υ_n of variable $\mathbf{x}_n$ are easily obtainable, they are not needed because this normal density integrates to unity regardless of their values. Since the $\mathbf{y}_{n-1}$ term is independent of $\mathbf{x}_n$, the state density becomes

$$\begin{aligned} \gamma(\mathbf{x}_{n-1}) &= \mathcal{N}(\mathbf{y}_{n-1}; \omega_n, \Omega_n) \\ &= \mathcal{N}\left(\mathbf{x}_{n-1}; \mu_{n-1|N}, \mathbf{P}_{n-1|N}\right). \end{aligned} \tag{124}$$

The recursions for calculating the conditional mean and covariance of each state from those corresponding to the next later state are respectively given by

$$\begin{aligned} \mu_{n-1|N} &= \omega_n + (\mathbf{I} - \mathbf{H}_n \mathbf{A}_n)\, \mu_{n-1|n-1} \\ &= \mathbf{H}_n \mu_{n|N} + \mu_{n-1|n-1} - \mathbf{H}_n \mu_{n|n-1} \\ &= \mu_{n-1|n-1} + \mathbf{H}_n \left(\mu_{n|N} - \mu_{n|n-1}\right) \end{aligned} \tag{125}$$

and

$$\begin{aligned} \mathbf{P}_{n-1|N} &= \Lambda_n + \mathbf{H}_n \mathbf{P}_{n|N} \mathbf{H}_n^{\mathrm{T}} \\ &= (\mathbf{I} - \mathbf{H}_n \mathbf{A}_n)\, \mathbf{P}_{n-1|n-1} + \mathbf{H}_n \mathbf{P}_{n|N} \mathbf{H}_n^{\mathrm{T}} \\ &= \mathbf{P}_{n-1|n-1} - \mathbf{H}_n \mathbf{A}_n \mathbf{P}_{n-1|n-1} + \mathbf{H}_n \mathbf{P}_{n|N} \mathbf{H}_n^{\mathrm{T}} \\ &= \mathbf{P}_{n-1|n-1} - \mathbf{H}_n \mathbf{P}_{n|n-1} \mathbf{H}_n^{\mathrm{T}} + \mathbf{H}_n \mathbf{P}_{n|N} \mathbf{H}_n^{\mathrm{T}} \\ &= \mathbf{P}_{n-1|n-1} + \mathbf{H}_n \left(\mathbf{P}_{n|N} - \mathbf{P}_{n|n-1}\right) \mathbf{H}_n^{\mathrm{T}}. \end{aligned} \tag{126}$$

These expressions equal the mean and covariance from the RTS formulation of the fixed-interval Kalman smoother [41]. This joint-density marginalization approach thus provides a very natural way of deriving the RTS smoothing algorithm.

3.3 VITERBI ALGORITHM

The forward passes of the Viterbi and Baum algorithms differ only in that the Viterbi algorithm maximizes over the previous state at each step whereas the Baum algorithm marginalizes out the previous state. For Gaussian state densities, the difference between marginalization and maximization is just a scale factor. Neglecting this scale factor, the Viterbi forward density functions are

$$\phi(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n; \mu_{n|n}, \mathbf{P}_{n|n}\right), \tag{127}$$

where $\mu_{n|n}$ and $\mathbf{P}_{n|n}$ are defined in equations (76) and (77), respectively.

The backward pass of the Viterbi algorithm begins by maximizing the forward density function at the terminal time step, which gives $\hat{\mathbf{x}}_N = \mu_{N|N}$ by inspection. After this notation is adopted for the nth state estimate (i.e., $\hat{\mathbf{x}}_n = \mu_{n|N}$), the function that is maximized during the Viterbi backward recursion can be written as

$$p(\mu_{n|N}|\mathbf{x}_{n-1})\,\phi(\mathbf{x}_{n-1}) \;=\; \mathcal{N}\Big(\mu_{n|N};\mathbf{A}_n\mathbf{x}_{n-1},\mathbf{Q}_n\Big)\,\mathcal{N}\Big(\mathbf{x}_{n-1};\mu_{n-1|n-1},\mathbf{P}_{n-1|n-1}\Big)\,. \quad (128)$$

Here, the term $p(\mathbf{z}_n|\mathbf{x}_n = \mu_{n|N})$ is neglected since it is a constant that does not affect the arg max operation. Applying the GRL and again neglecting a constant term yields

$$\hat{\mathbf{x}}_{n-1|N} \;=\; \arg\max_{\mathbf{x}_{n-1}}\Big\{\mathcal{N}\Big(\mathbf{x}_{n-1};\mu_{n-1|N},\mathbf{\Lambda}_n\Big)\Big\},$$

where $\mathbf{\Lambda}_n$ is defined in equation (71) and

$$\mu_{n-1|N} \;=\; \mu_{n-1|n-1} + \mathbf{H}_n\Big(\mu_{n|N} - \mu_{n|n-1}\Big)\,. \quad (129)$$

In this last expression, $\mu_{n|n-1} = \mathbf{A}_n\mu_{n-1|n-1}$ and $\mathbf{H}_n$ is defined in equation (72). The state estimate is $\hat{\mathbf{x}}_{n-1} = \mu_{n-1|N}$, which equals the smoothed state estimate from the RTS smoother [41]. Note, however, that $\mathbf{\Lambda}_n$ is *not* the covariance matrix for the state estimate. The Viterbi algorithm is, by its nature, incapable of providing second-order statistical information. On another note, the equivalence of the RTS and two-filter smoothers implies the equivalence of the most likely state sequence (the Viterbi track) and the sequence of individually most likely states (the Baum estimates).

3.4 PARAMETER ESTIMATION

For HGMMs, the E-step of the EM algorithm consists of evaluating the auxiliary-function components defined in equations (24),(25), and (26), which requires the calculation of expected values for various quadratic functions under a normal density. These expectations are evaluated using the following identity:

$$\int d\mathbf{x}\,\Big(\mathbf{x}^{\mathrm{T}}\mathbf{F}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{f} + f_0\Big)\,\mathcal{N}(\mathbf{x};\mu,\mathbf{P}) \;=\; \mathrm{tr}\Big\{\mathbf{F}\Big(\mathbf{P} + \mu\mu^{\mathrm{T}}\Big)\Big\} + \mu^{\mathrm{T}}\mathbf{f} + f_0, \quad (130)$$

which is a special case of theorem 10.5.1 in Graybill [46]. The M-step of the EM algorithm consists of maximizing the auxiliary-function components obtained during the E-step, which requires differentiation of those components. In addition to standard matrix and trace derivatives, these calculations require the identity [47]

$$\frac{\partial}{\partial\mathbf{F}}\mathrm{tr}\Big\{\mathbf{F}^{\mathrm{T}}\mathbf{S}_1\mathbf{F}\mathbf{S}_2\Big\} \;=\; \mathbf{S}_1\mathbf{F}\mathbf{S}_2 + \mathbf{S}_1^{\mathrm{T}}\mathbf{F}\mathbf{S}_2^{\mathrm{T}}. \quad (131)$$

Auxiliary-function component Q_0 is given by

$$\begin{aligned} Q_0 &= \sum_{k=1}^{K} \int d\mathbf{x}_0^k \, \gamma(\mathbf{x}_0^k) \log p(\mathbf{x}_0^k) \\ &= \sum_{k=1}^{K} \int d\mathbf{x}_0^k \, \mathcal{N}\left(\mathbf{x}_0^k; \mu_{0|N_k}^k, \mathbf{P}_{0|N_k}^k\right) \log \mathcal{N}\left(\mathbf{x}_0^k; \mu_0, \mathbf{P}_0\right). \end{aligned} \tag{132}$$

Expanding the logarithm term, neglecting the constant $-L\log(2\pi)/2$, and neglecting a scale factor of $1/2$ results in

$$Q_0 = \sum_{k=1}^{K} \int d\mathbf{x}_0^k \, \mathcal{N}\left(\mathbf{x}_0^k; \mu_{0|N_k}^k, \mathbf{P}_{0|N_k}^k\right) \left\{\log\left|\mathbf{P}_0^{-1}\right| - \left(\mathbf{x}_0^k - \mu_0\right)^{\mathrm{T}} \mathbf{P}_0^{-1} \left(\mathbf{x}_0^k - \mu_0\right)\right\}. \tag{133}$$

Performing the integration using equation (130) gives

$$Q_0 = K \log\left|\mathbf{P}_0^{-1}\right| - \sum_{k=1}^{K} \mathrm{tr}\left\{\mathbf{P}_0^{-1}\left[\mathbf{P}_{0|N_k}^k + \left(\mu_{0|N_k}^k - \mu_0\right)\left(\mu_{0|N_k}^k - \mu_0\right)^{\mathrm{T}}\right]\right\}. \tag{134}$$

This function is concave in the parameters μ_0 and $\mathbf{P}_0^{-1}$, so that the optimal parameter estimates occur at the unique critical point. The derivative of Q_0 with respect to μ_0 is obtained as

$$\begin{aligned} \frac{\partial Q_0}{\partial \mu_0} &= -\frac{\partial}{\partial \mu_0} \sum_{k=1}^{K} \mathrm{tr}\left\{\mathbf{P}_0^{-1}\left(\mu_{0|N_k}^k - \mu_0\right)\left(\mu_{0|N_k}^k - \mu_0\right)^{\mathrm{T}}\right\} \\ &= -\sum_{k=1}^{K} \frac{\partial}{\partial \mu_0}\left(\mu_{0|N_k}^k - \mu_0\right)^{\mathrm{T}} \mathbf{P}_0^{-1}\left(\mu_{0|N_k}^k - \mu_0\right) \\ &= -\sum_{k=1}^{K} 2\mathbf{P}_0^{-1}\left(\mu_{0|N_k}^k - \mu_0\right). \end{aligned} \tag{135}$$

Equating this derivative to zero and solving for μ_0 gives

$$\mu_0^{i+1} = \frac{1}{K} \sum_{k=1}^{K} \mu_{0|N_k}^k. \tag{136}$$

In general, a symmetry constraint must be imposed when maximizing Q_0 to find the optimal $\mathbf{P}_0$. This constraint can be implemented by using derivative formulas that explicitly take into account the symmetry of the matrix or by performing a constrained optimization in which the appropriate Lagrangian is maximized. The constraint turns out to be redundant for the HGMM covariance matrices, however, because the unconstrained optimizer has a symmetric form. This redundancy occurs when optimizing $\mathbf{Q}_n$ and $\mathbf{R}_n$ as well. The derivative expressions used to find the optimal covariance estimates are therefore given for matrices with independent elements, which greatly simplifies the analysis.

The derivative of equation (134) with respect to $\mathbf{P}_0^{-1}$ is

$$\frac{\partial Q_0}{\partial \mathbf{P}_0^{-1}} = K\,\mathbf{P}_0 - \sum_{k=1}^{K}\left\{\mathbf{P}_{0|N_k}^k + \left(\mu_{0|N_k}^k - \mu_0\right)\left(\mu_{0|N_k}^k - \mu_0\right)^{\mathrm{T}}\right\}. \tag{137}$$

Equating to zero, substituting the optimal value of μ_0, and defining

$$\varepsilon_0^k = \mu_{0|N_k}^k - \mu_0^{i+1} \tag{138}$$

gives the parameter update

$$\mathbf{P}_0^{i+1} = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathbf{P}_{0|N_k}^k + \varepsilon_0^k\,\varepsilon_0^{k\mathrm{T}}\right\}. \tag{139}$$

The transition-density component Q_X was defined in equation (25) as

$$Q_X = \sum_{n=1}^{N_{\max}}\sum_{k=1}^{K_n}\int d\mathbf{x}_n^k \int d\mathbf{x}_{n-1}^k\,\gamma(\mathbf{x}_n^k,\mathbf{x}_{n-1}^k)\,\log\,p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k).$$

Substituting the definition of $p(\mathbf{x}_n^k|\mathbf{x}_{n-1}^k)$, neglecting the constant $L\log(2\pi)/2$ and the scale factor $1/2$, and marginalizing $\gamma(\mathbf{x}_n^k,\mathbf{x}_{n-1}^k)$ from the $\log\left|\mathbf{Q}_n^{-1}\right|$ term yields

$$Q_X = \sum_{n=1}^{N_{\max}}\sum_{k=1}^{K_n}\left\{\log\left|\mathbf{Q}_n^{-1}\right| - I_k(\mathbf{A}_n,\mathbf{Q}_n)\right\}, \tag{140}$$

where

$$\begin{aligned} I_k(\mathbf{A}_n,\mathbf{Q}_n) &= \int d\mathbf{x}_n^k \int d\mathbf{x}_{n-1}^k\,\gamma(\mathbf{x}_n^k,\mathbf{x}_{n-1}^k) \\ &\quad\times\left\{\left(\mathbf{x}_n^k - \mathbf{A}_n\mathbf{x}_{n-1}^k\right)^{\mathrm{T}}\mathbf{Q}_n^{-1}\left(\mathbf{x}_n^k - \mathbf{A}_n\mathbf{x}_{n-1}^k\right)\right\}. \end{aligned} \tag{141}$$

The integral term can be further decomposed as

$$I_k(\mathbf{A}_n,\mathbf{Q}_n) = I_k^{(1)}(\mathbf{Q}_n) + I_k^{(2)}(\mathbf{A}_n,\mathbf{Q}_n) + I_k^{(3)}(\mathbf{A}_n,\mathbf{Q}_n). \tag{142}$$

The first component is given by

$$\begin{aligned} I_k^{(1)} &= \int d\mathbf{x}_n^k \int d\mathbf{x}_{n-1}^k\,\gamma(\mathbf{x}_n^k,\mathbf{x}_{n-1}^k)\left\{\mathbf{x}_n^{k\mathrm{T}}\,\mathbf{Q}_n^{-1}\,\mathbf{x}_n^k\right\} \\ &= \int d\mathbf{x}_n^k\,\gamma(\mathbf{x}_n^k)\,\mathbf{x}_n^{k\mathrm{T}}\,\mathbf{Q}_n^{-1}\,\mathbf{x}_n^k \\ &= \int d\mathbf{x}_n^k\,\mathcal{N}\left(\mathbf{x}_n^k;\mu_{n|N_k}^k,\mathbf{P}_{n|N_k}^k\right)\mathbf{x}_n^{k\mathrm{T}}\,\mathbf{Q}_n^{-1}\,\mathbf{x}_n^k \\ &= \mathrm{tr}\left\{\mathbf{Q}_n^{-1}\left(\mathbf{P}_{n|N_k}^k + \mu_{n|N_k}^k\,\mu_{n|N_k}^{k\mathrm{T}}\right)\right\}. \end{aligned} \tag{143}$$

The second component is

$$
\begin{aligned}
I_k^{(2)} &= -2\int d\mathbf{x}_n^k \int d\mathbf{x}_{n-1}^k\, \gamma(\mathbf{x}_n^k, \mathbf{x}_{n-1}^k)\left\{\mathbf{x}_n^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\mathbf{x}_{n-1}^k\right\} \\
&= -2\int d\mathbf{x}_n^k \int d\mathbf{x}_{n-1}^k\, \mathcal{N}\left(\mathbf{x}_n^k;\mu_{n|N_k}^k,\mathbf{P}_{n|N_k}^k\right)\mathcal{N}\left(\mathbf{x}_{n-1}^k;\lambda_n^k,\Lambda_n^k\right)\mathbf{x}_n^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\mathbf{x}_{n-1}^k \\
&= -2\int d\mathbf{x}_n^k\, \mathcal{N}\left(\mathbf{x}_n^k;\mu_{n|N_k}^k,\mathbf{P}_{n|N_k}^k\right)\mathbf{x}_n^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\int d\mathbf{x}_{n-1}^k\, \mathcal{N}\left(\mathbf{x}_{n-1}^k;\lambda_n^k,\Lambda_n^k\right)\mathbf{x}_{n-1}^k \\
&= -2\int d\mathbf{x}_n^k\, \mathcal{N}\left(\mathbf{x}_n^k;\mu_{n|N_k}^k,\mathbf{P}_{n|N_k}^k\right)\mathbf{x}_n^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\,\lambda_n^k \\
&= -2\int d\mathbf{x}_n^k\, \mathcal{N}\left(\mathbf{x}_n^k;\mu_{n|N_k}^k,\mathbf{P}_{n|N_k}^k\right)\mathbf{x}_n^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\left\{\left(\mathbf{I}-\mathbf{H}_n^k\mathbf{A}_n\right)\mu_{n-1|n-1}^k+\mathbf{H}_n^k\mathbf{x}_n^k\right\} \\
&= -2\int d\mathbf{x}_n^k\, \mathcal{N}\left(\mathbf{x}_n^k;\mu_{n|N_k}^k,\mathbf{P}_{n|N_k}^k\right) \\
&\qquad \times\left\{\mathbf{x}_n^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\mathbf{H}_n^k\mathbf{x}_n^k+\mathbf{x}_n^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\left(\mathbf{I}-\mathbf{H}_n^k\mathbf{A}_n\right)\mu_{n-1|n-1}^k\right\}. \qquad (144)
\end{aligned}
$$

Applying the identity in equation (130) then gives

$$
\begin{aligned}
I_k^{(2)} &= -2\,\mathrm{tr}\left\{\mathbf{Q}_n^{-1}\mathbf{A}_n\mathbf{H}_n^k\left(\mathbf{P}_{n|N_k}^k+\mu_{n|N_k}^k\,\mu_{n|N_k}^{k\mathrm{T}}\right)\right\}-2\,\mu_{n|N_k}^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\left(\mathbf{I}-\mathbf{H}_n^k\mathbf{A}_n\right)\mu_{n-1|n-1}^k \\
&= -2\,\mathrm{tr}\left\{\mathbf{Q}_n^{-1}\mathbf{A}_n\left(\mathbf{H}_n^k\,\mathbf{P}_{n|N_k}^k\right)\right\}-2\,\mu_{n|N_k}^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\left\{\mu_{n-1|n-1}^k+\mathbf{H}_n^k\left(\mu_{n|N_k}^k-\mu_{n|n-1}^k\right)\right\} \\
&= -2\,\mathrm{tr}\left\{\mathbf{Q}_n^{-1}\mathbf{A}_n\left(\mathbf{H}_n^k\,\mathbf{P}_{n|N_k}^k\right)\right\}-2\,\mu_{n|N_k}^{k\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\,\mu_{n-1|N_k}^k \\
&= -2\,\mathrm{tr}\left\{\mathbf{Q}_n^{-1}\mathbf{A}_n\left(\mathbf{P}_{n,n-1|N}^k+\mu_{n|N_k}^k\,\mu_{n-1|N_k}^{k\mathrm{T}}\right)^{\mathrm{T}}\right\}. \qquad (145)
\end{aligned}
$$

The last component is given by

$$
\begin{aligned}
I_k^{(3)}(\mathbf{A}_n,\mathbf{Q}_n) &= \int d\mathbf{x}_n^k \int d\mathbf{x}_{n-1}^k\, \gamma(\mathbf{x}_n^k,\mathbf{x}_{n-1}^k)\left\{\mathbf{x}_{n-1}^{k\mathrm{T}}\mathbf{A}_n^{\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\mathbf{x}_{n-1}^k\right\} \\
&= \int d\mathbf{x}_{n-2}^k\, \gamma(\mathbf{x}_{n-1}^k)\,\mathbf{x}_{n-1}^{k\mathrm{T}}\mathbf{A}_n^{\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\mathbf{x}_{n-1}^k \\
&= \int d\mathbf{x}_n^k\, \mathcal{N}\left(\mathbf{x}_{n-1}^k;\mu_{n-1|N_k}^k,\mathbf{P}_{n-1|N_k}^k\right)\mathbf{x}_{n-1}^{k\mathrm{T}}\mathbf{A}_n^{\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\mathbf{x}_{n-1}^k \\
&= \mathrm{tr}\left\{\mathbf{A}_n^{\mathrm{T}}\mathbf{Q}_n^{-1}\mathbf{A}_n\left(\mathbf{P}_{n-1|N_k}^k+\mu_{n-1|N_k}^k\,\mu_{n-1|N_k}^{k\mathrm{T}}\right)\right\} \\
&= \mathrm{tr}\left\{\mathbf{Q}_n^{-1}\mathbf{A}_n\left(\mathbf{P}_{n-1|N_k}^k+\mu_{n-1|N_k}^k\,\mu_{n-1|N_k}^{k\mathrm{T}}\right)\mathbf{A}_n^{\mathrm{T}}\right\}. \qquad (146)
\end{aligned}
$$

When these integral terms are substituted back into the expression for Q_X, each trace term undergoes a summation over k. Since the parameters $\mathbf{A}_n$ and $\mathbf{Q}_n$ are constant across k, the summations can be taken inside the trace and applied to the terms involving the state means and covariances. It is therefore convenient to define

$$
\mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} = \sum_{k=1}^{K_n}\left\{\mathbf{P}_{n|N_k}^k+\mu_{n|N_k}^k\mu_{n|N_k}^{k\mathrm{T}}\right\} \qquad (147)
$$

$$
\mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} = \sum_{k=1}^{K_n}\left\{\mathbf{P}_{n,n-1|N_k}^k+\mu_{n|N_k}^k\mu_{n-1|N_k}^{k\mathrm{T}}\right\}. \qquad (148)
$$

With these in hand, Q_X can be written as

$$\begin{aligned} Q_X &= \sum_{n=1}^{N_{\max}} \left\{ K_n \log \left| \mathbf{Q}_n^{-1} \right| - \operatorname{tr}\left\{ \mathbf{Q}_n^{-1} \left(\mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} - 2\,\mathbf{A}_n \mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}} + \mathbf{A}_n \mathbf{C}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}} \mathbf{A}_n^{\mathrm{T}} \right) \right\} \right\} \\ &= \sum_{n=1}^{N_{\max}} \left\{ K_n \log \left| \mathbf{Q}_n^{-1} \right| - \operatorname{tr}\left\{ \mathbf{Q}_n^{-1}\, \mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} \right\} + \operatorname{tr}\left\{ \mathbf{Q}_n^{-1}\, \mathbf{A}_n \mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}} \right\} \right. \\ &\qquad \left. + \operatorname{tr}\left\{ \mathbf{Q}_n^{-1} \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} \mathbf{A}_n^{\mathrm{T}} \right\} - \operatorname{tr}\left\{ \mathbf{Q}_n^{-1} \mathbf{A}_n \mathbf{C}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}} \mathbf{A}_n^{\mathrm{T}} \right\} \right\}. \end{aligned} \tag{149}$$

This last expression, which makes use of the trace properties, is used for optimization of the covariance matrix because it ensures symmetry in cases when $\mathbf{A}_n$ is known [48]. The derivative of Q_X with respect to $\mathbf{A}_n$ is

$$\begin{aligned} \frac{\partial Q_X}{\partial \mathbf{A}_n} &= -\frac{\partial}{\partial \mathbf{A}_n} \operatorname{tr}\left\{ \mathbf{Q}_n^{-1} \left(\mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} - 2\,\mathbf{A}_n \mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}} + \mathbf{A}_n \mathbf{C}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}} \mathbf{A}_n^{\mathrm{T}} \right) \right\} \\ &= 2\frac{\partial}{\partial \mathbf{A}_n} \operatorname{tr}\left\{ \mathbf{A}_n \mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}} \mathbf{Q}_n^{-1} \right\} - 2\frac{\partial}{\partial \mathbf{A}_n} \operatorname{tr}\left\{ \mathbf{A}_n^{\mathrm{T}} \mathbf{Q}_n^{-1} \mathbf{A}_n \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} \right\} \\ &= 2\left(\mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}} \mathbf{Q}_n^{-1} \right)^{\mathrm{T}} - 2\,\mathbf{Q}_n^{-1} \mathbf{A}_n \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} \\ &= 2\,\mathbf{Q}_n^{-1} \left(\mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} - \mathbf{A}_n \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} \right). \end{aligned} \tag{150}$$

Equating this derivative to zero and solving for $\mathbf{A}_n$ yields the update

$$\mathbf{A}_n^{i+1} = \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} \mathbf{C}^{-1}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}}. \tag{151}$$

The derivative of Q_X with respect to $\mathbf{Q}_n^{-1}$ is

$$\frac{\partial Q_X}{\partial \mathbf{Q}_n^{-1}} = K_n \mathbf{Q}_n - \mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} + \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} \mathbf{A}_n^{\mathrm{T}} + \mathbf{A}_n \mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}} - \mathbf{A}_n \mathbf{C}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}} \mathbf{A}_n^{\mathrm{T}}, \tag{152}$$

which, when equated to zero, gives the estimator

$$\mathbf{Q}_n^{i+1} = \frac{1}{K_n} \left\{ \mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} - \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} \mathbf{A}_n^{\mathrm{T}} - \mathbf{A}_n \mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}} + \mathbf{A}_n \mathbf{C}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}} \mathbf{A}_n^{\mathrm{T}} \right\}. \tag{153}$$

This expression is used if $\mathbf{A}_n$ is known. Otherwise, the optimal value for $\mathbf{A}_n$ is substituted to obtain the update

$$\mathbf{Q}_n^{i+1} = \frac{1}{K_n} \left\{ \mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} - \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} \mathbf{C}^{-1}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}} \mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}} \right\}. \tag{154}$$

Finally, the auxiliary-function component Q_Z is

$$\begin{aligned} Q_Z &= \sum_{n=1}^{N_{\max}} \sum_{k=1}^{K_n} \int d\mathbf{x}_n^k \; \gamma(\mathbf{x}_n^k) \log p(\mathbf{z}_n^k | \mathbf{x}_n^k) \\ &= \sum_{n=1}^{N_{\max}} \sum_{k=1}^{K_n} \left\{ \log \left| \mathbf{R}_n^{-1} \right| - I_k(\mathbf{B}_n, \mathbf{R}_n) \right\}, \end{aligned} \tag{155}$$

where the integral term is

$$
\begin{aligned}
I_k(\mathbf{B}_n, \mathbf{R}_n) &= \int d\mathbf{x}_n^k \, \mathcal{N}\left(\mathbf{x}_n^k; \mu_{n|N_k}^k, \mathbf{P}_{n|N_k}^k\right) \left\{ \left(\mathbf{z}_n^k - \mathbf{B}_n \mathbf{x}_n^k\right)^{\mathrm{T}} \mathbf{R}_n^{-1} \left(\mathbf{z}_n^k - \mathbf{B}_n \mathbf{x}_n^k\right) \right\} \\
&= \int d\mathbf{x}_n^k \, \mathcal{N}\left(\mathbf{x}_n^k; \mu_{n|N_k}^k, \mathbf{P}_{n|N_k}^k\right) \\
&\qquad \times \left\{ \mathbf{z}_n^{k\mathrm{T}} \mathbf{R}_n^{-1} \mathbf{z}_n^k - 2\, \mathbf{x}_n^{k\mathrm{T}} \mathbf{B}_n^{\mathrm{T}} \mathbf{R}_n^{-1} \mathbf{z}_n^k + \mathbf{x}_n^{k\mathrm{T}} \mathbf{B}_n^{\mathrm{T}} \mathbf{R}_n^{-1} \mathbf{B}_n \mathbf{x}_n^k \right\} \\
&= \mathbf{z}_n^{k\mathrm{T}} \mathbf{R}_n^{-1} \mathbf{z}_n^k - 2\, \mu_{n|N_k}^{k\mathrm{T}} \mathbf{B}_n^{\mathrm{T}} \mathbf{R}_n^{-1} \mathbf{z}_n^k + \mathrm{tr}\left\{ \mathbf{B}_n^{\mathrm{T}} \mathbf{R}_n^{-1} \mathbf{B}_n \left(\mathbf{P}_{n|N_k}^k + \mu_{n|N_k}^k \, \mu_{n|N_k}^{k\mathrm{T}} \right) \right\} \\
&= \mathrm{tr}\left\{ \mathbf{R}_n^{-1} \left\{ \mathbf{z}_n^k \, \mathbf{z}_n^{k\mathrm{T}} - 2\, \mathbf{z}_n^k \, \mu_{n|N_k}^{k\mathrm{T}} \mathbf{B}_n^{\mathrm{T}} + \mathbf{B}_n \left(\mathbf{P}_{n|N_k}^k + \mu_{n|N_k}^k \, \mu_{n|N_k}^{k\mathrm{T}} \right) \mathbf{B}_n^{\mathrm{T}} \right\} \right\}.
\end{aligned}
\tag{156}
$$

After bringing the summation over k into the trace terms and defining

$$
\mathbf{C}_{\mathbf{z}_n \mathbf{z}_n} = \sum_{k=1}^{K_n} \mathbf{z}_n^k \, \mathbf{z}_n^{k\mathrm{T}} \tag{157}
$$

$$
\mathbf{C}_{\mathbf{z}_n \mathbf{x}_n} = \sum_{k=1}^{K_n} \mathbf{z}_n^k \, \mu_{n|N_k}^{k\mathrm{T}}, \tag{158}
$$

Q_Z can be expressed as

$$
\begin{aligned}
Q_Z &= \sum_{n=1}^{N_{\max}} \left\{ K_n \log \left| \mathbf{R}_n^{-1} \right| - \mathrm{tr}\left\{ \mathbf{R}_n^{-1} \left(\mathbf{C}_{\mathbf{z}_n \mathbf{z}_n} - 2\, \mathbf{B}_n \mathbf{C}_{\mathbf{z}_n \mathbf{x}_n}^{\mathrm{T}} + \mathbf{B}_n \mathbf{C}_{\mathbf{x}_n \mathbf{x}_n} \mathbf{B}_n^{\mathrm{T}} \right) \right\} \right\} \\
&= \sum_{n=1}^{N_{\max}} \left\{ K_n \log \left| \mathbf{R}_n^{-1} \right| - \mathrm{tr}\left\{ \mathbf{R}_n^{-1} \, \mathbf{C}_{\mathbf{z}_n \mathbf{z}_n} \right\} + \mathrm{tr}\left\{ \mathbf{R}_n^{-1} \, \mathbf{B}_n \mathbf{C}_{\mathbf{z}_n \mathbf{x}_n}^{\mathrm{T}} \right\} \right. \\
&\qquad \left. + \mathrm{tr}\left\{ \mathbf{R}_n^{-1} \mathbf{C}_{\mathbf{z}_n \mathbf{x}_n} \mathbf{B}_n^{\mathrm{T}} \right\} - \mathrm{tr}\left\{ \mathbf{R}_n^{-1} \mathbf{B}_n \mathbf{C}_{\mathbf{x}_n \mathbf{x}_n} \mathbf{B}_n^{\mathrm{T}} \right\} \right\}.
\end{aligned}
\tag{159}
$$

Because this expression has the same structure as Q_X, the optimal updates are found at the critical point using the same steps as for Q_X, with the following result:

$$
\mathbf{B}_n^{i+1} = \mathbf{C}_{\mathbf{z}_n \mathbf{x}_n} \mathbf{C}_{\mathbf{x}_n \mathbf{x}_n}^{-1} \tag{160}
$$

$$
\mathbf{R}_n^{i+1} = \frac{1}{K_n} \left\{ \mathbf{C}_{\mathbf{z}_n \mathbf{z}_n} - \mathbf{C}_{\mathbf{z}_n \mathbf{x}_n} \mathbf{B}_n^{\mathrm{T}} - \mathbf{B}_n \mathbf{C}_{\mathbf{z}_n \mathbf{x}_n}^{\mathrm{T}} + \mathbf{B}_n \mathbf{C}_{\mathbf{x}_n \mathbf{x}_n} \mathbf{B}_n^{\mathrm{T}} \right\} \tag{161}
$$

$$
= \frac{1}{K_n} \left\{ \mathbf{C}_{\mathbf{z}_n \mathbf{z}_n} - \mathbf{C}_{\mathbf{z}_n \mathbf{x}_n} \mathbf{C}_{\mathbf{x}_n \mathbf{x}_n}^{-1} \mathbf{C}_{\mathbf{z}_n \mathbf{x}_n}^{\mathrm{T}} \right\}. \tag{162}
$$

The correlation matrix and model parameter estimates are shown in table 8. In this table, the factors $1/K_n$ on the correlation matrices are borrowed from $\mathbf{Q}_n^{i+1}$ and $\mathbf{R}_n^{i+1}$. The factors cancel in $\mathbf{A}_n^{i+1}$ and $\mathbf{B}_n^{i+1}$.

Table 8. EM Parameter Estimators for HGMMs

Correlation Matrices	$\mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} = \frac{1}{K_n}\sum_{k=1}^{K_n}\left\{\mathbf{P}^k_{n\|N_k} + \mu^k_{n\|N_k}\,\mu^{k\mathrm{T}}_{n\|N_k}\right\}$ $\mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} = \frac{1}{K_n}\sum_{k=1}^{K_n}\left\{\mathbf{P}^k_{n,n-1\|N_k} + \mu^k_{n\|N_k}\,\mu^{k\mathrm{T}}_{n-1\|N_k}\right\}$ $\mathbf{C}_{\mathbf{z}_n\mathbf{z}_n} = \frac{1}{K_n}\sum_{k=1}^{K_n}\mathbf{z}^k_n\,\mathbf{z}^{k\mathrm{T}}_n$ $\mathbf{C}_{\mathbf{z}_n\mathbf{x}_n} = \frac{1}{K_n}\sum_{k=1}^{K_n}\mathbf{z}^k_n\,\mu^{k\mathrm{T}}_{n\|N_k}$
System Matrices	$\mathbf{A}^{i+1}_n = \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}}\mathbf{C}^{-1}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}}$ $\mathbf{B}^{i+1}_n = \mathbf{C}_{\mathbf{z}_n\mathbf{x}_n}\mathbf{C}^{-1}_{\mathbf{x}_n\mathbf{x}_n}$ $\mathbf{Q}^{i+1}_n = \mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} - \mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}}\mathbf{C}^{-1}_{\mathbf{x}_{n-1}\mathbf{x}_{n-1}}\mathbf{C}^{\mathrm{T}}_{\mathbf{x}_n\mathbf{x}_{n-1}}$ $\mathbf{R}^{i+1}_n = \mathbf{C}_{\mathbf{z}_n\mathbf{z}_n} - \mathbf{C}_{\mathbf{z}_n\mathbf{x}_n}\mathbf{C}^{-1}_{\mathbf{x}_n\mathbf{x}_n}\mathbf{C}^{\mathrm{T}}_{\mathbf{z}_n\mathbf{x}_n}$
Initial-State Parameters	$\mu^{i+1}_0 = \frac{1}{K}\sum_{k=1}^{K}\mu^k_{0\|N_k}$ $\varepsilon^k_0 = \mu^k_{0\|N_k} - \mu^{i+1}_0$ $\mathbf{P}^{i+1}_0 = \frac{1}{K}\sum_{k=1}^{K}\left\{\mathbf{P}^k_{0\|N_k} + \varepsilon^k_0\,\varepsilon^{k\mathrm{T}}_0\right\}$

The auxiliary-function components Q_0, Q_X, and Q_Z are concave in parameter sets $\left\{\mu_0, \mathbf{P}_0^{-1}\right\}$, $\{\mathbf{A}_n, \mathbf{Q}_n^{-1}\}$, and $\{\mathbf{B}_n, \mathbf{R}_n^{-1}\}$, respectively. The updates at each iteration are therefore the unique maxima of the CDLF. The final measurement likelihood $p(\mathbf{Z}_N)$ is not necessarily concave, however, so that suboptimal local maxima are possible. Multiple training runs from different starting points may therefore be needed to find the global maximum.

3.5 TIME-INVARIANT MODELS

The parameter update formulas given above are easily specialized to time-invariant HGMMs by performing a second averaging operation across time when the correlation matrices are calculated in equations (147), (148), (157), and (158). These time-averaged correlation matrices are then used in equations (151), (154), (160), and (162), which are each evaluated only once for all time.

The measurement likelihood in the time-invariant case has a parameter invariance structure that is worth noting. Specifically, as an argument of the measurement likelihood function in equation (82), the parameter set

$$\Theta = \{\mathbf{A}, \mathbf{B}, \mathbf{Q}, \mathbf{R}, \mu_0, \mathbf{P}_0\} \tag{163}$$

is equivalent to any set

$$\tilde{\Theta} = \left\{\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{Q}}, \mathbf{R}, \tilde{\mu}_0, \tilde{\mathbf{P}}_0\right\}, \tag{164}$$

where the transformed parameters are given by

$$\tilde{\mathbf{A}} = \mathbf{U}_A \mathbf{A} \mathbf{U}_A^{-1} \tag{165}$$

$$\tilde{\mathbf{B}} = \mathbf{B} \mathbf{U}_0^{-1} \mathbf{U}_A^{-1} \tag{166}$$

$$\tilde{\mathbf{Q}} = \mathbf{U}_A \mathbf{U}_0 \mathbf{Q} \mathbf{U}_0^{\mathrm{T}} \mathbf{U}_A^{\mathrm{T}} \tag{167}$$

$$\tilde{\mu}_0 = \mathbf{U}_A \mathbf{U}_0 \mu_0 \tag{168}$$

$$\tilde{\mathbf{P}}_0 = \mathbf{U}_A \mathbf{U}_0 \mathbf{P}_0 \mathbf{U}_0^{\mathrm{T}} \mathbf{U}_A^{\mathrm{T}}. \tag{169}$$

Here, $\mathbf{U}_A$ is any nonsingular $L \times L$ matrix, and $\mathbf{U}_0$ is any nonsingular $L \times L$ matrix that commutes with $\mathbf{A}$.

The equivalence of the two models under the likelihood function is demonstrated in appendix C. Two conclusions can be drawn concerning this invariance. First, the EM algorithm will converge to the member of the invariant family that is closest to the starting point. Second, any member of the invariant family is theoretically as good as any other for classification. While the second conclusion suggests that there is no need to be concerned about the first, it may be desirable for numerical reasons to constrain the EM algorithm to produce estimates of a given structure. For example, it might be beneficial for the transition matrix to be as close as possible to an identity matrix.

3.6 MIXED-MODE MODELS

Attention is now turned to mixed-mode HGMMs (MM-HGMMs). The Baum forward densities for these models are obtained directly by substituting an indexed version of equation (81) into equation (29), giving

$$\alpha(\mathbf{x}_n) \;=\; \sum_{j=1}^{J} \rho_j \, c_n^j \, \mathcal{N}\!\left(\mathbf{x}_n; \mu_{n|n}^j, \mathbf{P}_{n|n}^j\right). \tag{170}$$

The c_n^j, $\mu_{n|n}^j$, and $\mathbf{P}_{n|n}^j$ are calculated for each j using the HGMM recursions with the single-mode prior $p(\mathbf{x}_0|j)$. Furthermore, since the integral of the sum of terminal densities is just the sum of the integrals of the individual densities, the likelihood can be written as

$$p(\mathbf{Z}_N) = \sum_{j=1}^{J} \rho_j \, p(\mathbf{Z}_N|j) = \sum_{j=1}^{J} \rho_j \, c_N^j. \tag{171}$$

Drawing on equations (107) and (34), the conditional state densities are

$$\gamma(\mathbf{x}_n) \;=\; \sum_{j=1}^{J} \rho_{j|N} \, \mathcal{N}\!\left(\mathbf{x}_n; \mu_{n|N}^j, \mathbf{P}_{n|N}^j\right), \tag{172}$$

where $\mu_{n|N}^j$ is the conditional state mean and $\mathbf{P}_{n|N}^j$ is the state covariance matrix from the appropriate single-mode HGMM.

Parameter estimates in MM-HGMMs are obtained using results from section 2 and the analysis techniques of the previous subsection. The mean and covariance updates for the mixture components in the initial-state prior distribution are given by

$$\mu_{0,j}^{i+1} = \frac{1}{\kappa_j} \sum_{k=1}^{K} \rho_{j|N_k} \, \mu_{0|N_k}^{jk} \tag{173}$$

$$\mathbf{P}_{0,j}^{i+1} = \frac{1}{\kappa_j} \sum_{k=1}^{K} \rho_{j|N_k} \left\{ \mathbf{P}_{0|N_k}^{jk} + \varepsilon_0^{jk,i+1} \, \varepsilon_0^{jk,i+1\mathrm{T}} \right\}, \tag{174}$$

where

$$\kappa_j = \sum_{k=1}^{K} \rho_{j|N_k}. \tag{175}$$

The variable $\varepsilon_0^{jk,i+1}$ denotes the difference between the estimate of the mean at time t_0, conditioned on the kth measurement sequence, and the weighted sum of the estimates from all measurement sequences; that is,

$$\varepsilon_0^{jk,i+1} = \mu_{0|N_k}^{jk} - \mu_{0,j}^{i+1}. \tag{176}$$

The estimates for the mixing parameters were defined in section 2. The expressions for the system-matrix estimators in terms of correlation matrices are identical to those given in equations (151), (154), (160), and (162) for single-mode HGMMs. The correlation matrices for MM-HGMMs are very similar to those for single-mode HGMMs, but with a weighted sum over the mode assignments; that is,

$$\mathbf{C}_{\mathbf{x}_n\mathbf{x}_n} = \sum_{j=1}^{J} \rho_{j|N_k}\, \mathbf{C}^j_{\mathbf{x}_n\mathbf{x}_n} \tag{177}$$

$$\mathbf{C}_{\mathbf{x}_n\mathbf{x}_{n-1}} = \sum_{j=1}^{J} \rho_{j|N_k}\, \mathbf{C}^j_{\mathbf{x}_n\mathbf{x}_{n-1}} \tag{178}$$

$$\mathbf{C}_{\mathbf{z}_n\mathbf{x}_n} = \sum_{j=1}^{J} \rho_{j|N_k}\, \mathbf{C}^j_{\mathbf{z}_n\mathbf{x}_n}, \tag{179}$$

where $\mathbf{C}^j_{\mathbf{x}_n\mathbf{x}_n}$, $\mathbf{C}^j_{\mathbf{x}_n\mathbf{x}_{n-1}}$, and $\mathbf{C}^j_{\mathbf{z}_n\mathbf{x}_n}$ are given by equations (147), (148), and (158), respectively, except that the $\mu^k_{n|N_k}$ and $\mathbf{P}^k_{n|N_k}$ are indexed by j. The measurement correlation matrix is identical to that given in equation (157) since the measurements do not depend on the mode index.

While the number of components, J, in the mixture density has been assumed to be known thus far, it might be estimated as follows. First, the system matrices for a single-mode HGMM with a fixed large-variance Gaussian prior are estimated using the measurement training sequences. Second, the estimated parameters are used to calculate the conditional state sequence corresponding to each measurement sequence. Finally, a multivariate clustering algorithm is applied to the initial states from these state-sequence estimates. The number of significant clusters provides an estimate for J, and the location and spread of the clusters provide initial estimates of $\mu_{0,j}$ and $\mathbf{P}_{0,j}$, respectively, for each mixture component.

4. SUMMARY

A general theory of continuous-state hidden Markov models (CS-HMMs) has been presented. The given results solve the likelihood evaluation, state estimation, and parameter estimation problems, to the extent that the solutions can be formulated independent of the particular form of the model densities. The CS-HMM results were then specialized to linear Gaussian models, resulting in the hidden Gauss-Markov model (HGMM). A Gaussian refactorization lemma has been derived, which provides a necessary tool for evaluation of the Baum recursions for HGMMs and, at the same time, naturally generates the update recursions for Kalman filters and smoothers. The Baum and Viterbi algorithms for HGMMs were shown to be equivalent to two different implementations of the fixed-interval Kalman-smoother. It was shown that the likelihood obtained using the Baum algorithm for HGMMs equals the classical likelihood definition from Kalman-filter theory and that the parameter estimation algorithm for these models is equivalent to the existing expectation-maximization (EM) algorithm for Kalman-filter models. Taken together, these results unify Kalman-filter and HMM theory. The parameter-estimation algorithms given in this report were formulated for multiple training sequences with unequal lengths. The HGMM training algorithm presented here therefore extends previous algorithms that treat equal-length training measurements. This analysis has also resulted in a new estimator for the cross covariance between adjacent HGMM states that is considerably simpler than existing estimators. A parameter invariance structure was demonstrated for HGMMs whose parameters do not vary with time. Finally, the CS-HMM and HGMM algorithms were extended for models whose initial state is governed by mixtures of densities instead of a single density.

This work paves the way for extensions of HMM and Kalman-filter algorithms in both classification and tracking applications. For example, it generates a framework for investigating analogs of Kalman filters and smoothers for non-Gaussian model densities. In addition, the parameter-estimation algorithm could be used to obtain a more sophisticated, and possibly more accurate, tracking algorithm whose state and measurement covariance matrices are influenced by the observed data instead of being predetermined from the prior estimates of the model parameters.

APPENDIX A
PROOF OF THE REFACTORIZATION LEMMA

This appendix demonstrates the equivalence of the two product functions

$$\eta(\mathbf{y},\mathbf{x}) = \mathcal{N}(\mathbf{y};\mathbf{F}\mathbf{x},\mathbf{S})\,\mathcal{N}(\mathbf{x};\mu,\mathbf{P}) \tag{180}$$

$$= \mathcal{N}(\mathbf{y};\omega,\Omega)\,\mathcal{N}(\mathbf{x};\lambda,\Lambda)\,, \tag{181}$$

where ω, Ω, λ, and Λ are defined in terms of $\mathbf{F}$, $\mathbf{S}$, μ, and $\mathbf{P}$, as given in section 3.1. Essentially, this refactorization is equivalent to saying that $p(\mathbf{y}|\mathbf{x})\,p(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})\,p(\mathbf{y})$.

Beginning with equation (180) and substituting the normal densities gives

$$\eta(\mathbf{y},\mathbf{x}) = (2\pi)^{-(L+M)/2}\left|\mathbf{S}\right|^{-1/2}\left|\mathbf{P}\right|^{-1/2}\exp\left\{-\frac{1}{2}\xi(\mathbf{y},\mathbf{x})\right\}. \tag{182}$$

A useful form for the exponent $\xi(\mathbf{y},\mathbf{x})$ is obtained as

$$\begin{aligned}
\xi(\mathbf{y},\mathbf{x}) &= (\mathbf{y}-\mathbf{F}\mathbf{x})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{y}-\mathbf{F}\mathbf{x}) + (\mathbf{x}-\mu)^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\mu) \\
&= \mathbf{x}^{\mathrm{T}}\left(\mathbf{P}^{-1}+\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{F}\right)\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\left(\mathbf{P}^{-1}\mu+\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}\right) + \mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} + \mu^{\mathrm{T}}\mathbf{P}^{-1}\mu \\
&= \mathbf{x}^{\mathrm{T}}\Lambda^{-1}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\Lambda^{-1}\lambda + \mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} + \mu^{\mathrm{T}}\mathbf{P}^{-1}\mu \\
&= (\mathbf{x}-\lambda)^{\mathrm{T}}\Lambda^{-1}(\mathbf{x}-\lambda) - \lambda^{\mathrm{T}}\Lambda^{-1}\lambda + \mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} + \mu^{\mathrm{T}}\mathbf{P}^{-1}\mu\,.
\end{aligned} \tag{183}$$

The variables Λ and λ, which are introduced in the third step of equation (183) and are designed to facilitate completing the square in the last step, are defined by

$$\Lambda^{-1} = \mathbf{P}^{-1}+\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{F} \tag{184}$$

$$\Lambda^{-1}\lambda = \mathbf{P}^{-1}\mu+\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}. \tag{185}$$

These variables naturally arise in the "information" form, where the inverse covariance matrix is the primary variable instead of the covariance matrix and the mean is scaled by the inverse covariance. The variable defined in equation (184) is called the *information matrix*, and the variable defined in equation (185) is called the *information vector*. Equations (184) and (185) constitute a measurement update in the information formulation of a Kalman filter [5]. If the covariance matrix is desired, the matrix inversion lemma (MIL) [46, 49] can be applied to obtain

$$\Lambda = \mathbf{P}-\mathbf{P}\mathbf{F}^{\mathrm{T}}\left(\mathbf{S}+\mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}}\right)^{-1}\mathbf{F}\mathbf{P}. \tag{186}$$

Defining the variables

$$\Sigma = \mathbf{S} + \mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}} \tag{187}$$

$$\mathbf{H} = \mathbf{P}\mathbf{F}^{\mathrm{T}}\Sigma^{-1} \tag{188}$$

results in the covariance matrix becoming

$$\begin{aligned} \Lambda &= \mathbf{P} - \mathbf{P}\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\mathbf{F}\mathbf{P} \\ &= (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mathbf{P}. \end{aligned} \tag{189}$$

The mean vector is then obtained as

$$\begin{aligned} \lambda &= \Lambda\left(\mathbf{P}^{-1}\mu + \mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}\right) \\ &= (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mu + (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mathbf{P}\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mu + \left(\mathbf{I} - \mathbf{P}\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\mathbf{F}\right)\mathbf{P}\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mu + \mathbf{P}\mathbf{F}^{\mathrm{T}}\left(\mathbf{I} - \Sigma^{-1}\mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}}\right)\mathbf{S}^{-1}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mu + \left(\mathbf{P}\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\right)\left(\Sigma - \mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}}\right)\mathbf{S}^{-1}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mu + \mathbf{H}\mathbf{S}\mathbf{S}^{-1}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H}\mathbf{F})\,\mu + \mathbf{H}\mathbf{y}. \end{aligned} \tag{190}$$

Returning to the exponent term in equation (183) and introducing the functional

$$\zeta(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} + \mu^{\mathrm{T}}\mathbf{P}^{-1}\mu - \lambda^{\mathrm{T}}\Lambda^{-1}\lambda \tag{191}$$

causes the exponent to become

$$\xi(\mathbf{y},\mathbf{x}) = (\mathbf{x} - \lambda)^{\mathrm{T}}\Lambda^{-1}(\mathbf{x} - \lambda) + \zeta(\mathbf{y}), \tag{192}$$

which is put in a form similar to equation (183) as

$$\begin{aligned} \zeta(\mathbf{y}) &= \mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} + \mu^{\mathrm{T}}\mathbf{P}^{-1}\mu - \left(\mathbf{P}^{-1}\mu + \mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}\right)^{\mathrm{T}}\Lambda\left(\mathbf{P}^{-1}\mu + \mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}\right) \\ &= \mathbf{y}^{\mathrm{T}}\left(\mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{F}\Lambda\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\right)\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{F}\Lambda\mathbf{P}^{-1}\mu + \mu^{\mathrm{T}}\left(\mathbf{P}^{-1} - \mathbf{P}^{-1}\Lambda\mathbf{P}^{-1}\right)\mu \\ &= \mathbf{y}^{\mathrm{T}}\Omega^{-1}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\Omega^{-1}\omega + \mu^{\mathrm{T}}\left(\mathbf{P}^{-1} - \mathbf{P}^{-1}\Lambda\mathbf{P}^{-1}\right)\mu \\ &= (\mathbf{y} - \omega)^{\mathrm{T}}\Omega^{-1}(\mathbf{y} - \omega) - \omega^{\mathrm{T}}\Omega^{-1}\omega + \mu^{\mathrm{T}}\left(\mathbf{P}^{-1} - \mathbf{P}^{-1}\Lambda\mathbf{P}^{-1}\right)\mu. \end{aligned} \tag{193}$$

The variables ω and Ω introduced above are defined by

$$\Omega^{-1} = \mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{F}\Lambda\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1} \tag{194}$$

$$\Omega^{-1}\omega = \mathbf{S}^{-1}\mathbf{F}\Lambda\mathbf{P}^{-1}\mu. \tag{195}$$

These terms can be simplified by introducing the variable

$$\begin{aligned} \mathbf{D} &= \mathbf{S}^{-1}\mathbf{F}\Lambda \\ &= \mathbf{S}^{-1}\mathbf{F}\left(\mathbf{P}^{-1}+\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{F}\right)^{-1}, \end{aligned} \tag{196}$$

with which the information matrix becomes

$$\begin{aligned} \Omega^{-1} &= \left(\mathbf{I}-\mathbf{S}^{-1}\mathbf{F}\Lambda\mathbf{F}^{\mathrm{T}}\right)\mathbf{S}^{-1} \\ &= \left(\mathbf{I}-\mathbf{D}\mathbf{F}^{\mathrm{T}}\right)\mathbf{S}^{-1} \end{aligned} \tag{197}$$

and the information vector becomes

$$\Omega^{-1}\omega = \mathbf{D}\mathbf{P}^{-1}\mu. \tag{198}$$

Equation (194) can also be converted to covariance form by applying the MIL in reverse; that is,

$$\begin{aligned} \Omega &= \left\{\mathbf{S}^{-1}-\mathbf{S}^{-1}\mathbf{F}\left(\mathbf{P}^{-1}+\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\right\}^{-1} \\ &= \mathbf{S}+\mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}} = \Sigma. \end{aligned} \tag{199}$$

Noting that $\Omega = \Sigma$ and applying the MIL allows the mean vector to be obtained as

$$\begin{aligned} \omega &= \Omega\mathbf{S}^{-1}\mathbf{F}\left(\mathbf{P}^{-1}+\mathbf{F}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{F}\right)^{-1}\mathbf{P}^{-1}\mu \\ &= \Sigma\mathbf{S}^{-1}\mathbf{F}\left(\mathbf{P}-\mathbf{P}\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\mathbf{F}\mathbf{P}\right)\mathbf{P}^{-1}\mu \\ &= \Sigma\mathbf{S}^{-1}\left(\mathbf{I}-\mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\right)\mathbf{F}\mathbf{P}\mathbf{P}^{-1}\mu \\ &= \Sigma\mathbf{S}^{-1}\left(\Sigma-\mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}}\right)\Sigma^{-1}\mathbf{F}\mu \\ &= \Sigma\mathbf{S}^{-1}\mathbf{S}\Sigma^{-1}\mathbf{F}\mu \\ &= \mathbf{F}\mu. \end{aligned} \tag{200}$$

The functional $\zeta(\mathbf{y})$ is now written as

$$\zeta(\mathbf{y}) = (\mathbf{y}-\omega)^{\mathrm{T}}\Omega^{-1}(\mathbf{y}-\omega)+\kappa, \tag{201}$$

where

$$\begin{aligned} \kappa &= \mu^{\mathrm{T}}\left(\mathbf{P}^{-1}-\mathbf{P}^{-1}\Lambda\mathbf{P}^{-1}\right)\mu-\omega^{\mathrm{T}}\Omega^{-1}\omega \\ &= \mu^{\mathrm{T}}\left(\mathbf{P}^{-1}-\mathbf{P}^{-1}\Lambda\mathbf{P}^{-1}-\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\mathbf{F}\right)\mu \\ &= \mu^{\mathrm{T}}\left\{\mathbf{P}^{-1}-\mathbf{P}^{-1}\left(\mathbf{P}-\mathbf{P}\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\mathbf{F}\mathbf{P}\right)\mathbf{P}^{-1}-\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\mathbf{F}\right\}\mu \\ &= \mu^{\mathrm{T}}\left(\mathbf{P}^{-1}-\mathbf{P}^{-1}+\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\mathbf{F}-\mathbf{F}^{\mathrm{T}}\Sigma^{-1}\mathbf{F}\right)\mu \\ &= \mathbf{0}. \end{aligned} \tag{202}$$

The combined exponent from equation (180) therefore becomes

$$\xi(\mathbf{y},\mathbf{x}) = (\mathbf{y}-\omega)^{\mathrm{T}}\,\boldsymbol{\Omega}^{-1}\,(\mathbf{y}-\omega) + (\mathbf{x}-\lambda)^{\mathrm{T}}\,\boldsymbol{\Lambda}^{-1}\,(\mathbf{x}-\lambda).$$

Substituting this expression for $\xi(\mathbf{y},\mathbf{x})$ into equation (182) gives

$$\begin{aligned}\eta(\mathbf{y},\mathbf{x}) &= (2\pi)^{-(L+M)/2}\left|\mathbf{S}\right|^{-1/2}\left|\mathbf{P}\right|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\omega)^{\mathrm{T}}\,\boldsymbol{\Omega}^{-1}\,(\mathbf{y}-\omega)\right\} \\ &\quad\times\ \exp\left\{-\frac{1}{2}(\mathbf{x}-\lambda)^{\mathrm{T}}\,\boldsymbol{\Lambda}^{-1}\,(\mathbf{x}-\lambda)\right\}. \qquad (203)\end{aligned}$$

The exponentials can be written as normal densities if constants are included to compensate for the normalizing constants; that is,

$$\begin{aligned}\eta(\mathbf{y},\mathbf{x}) &= (2\pi)^{-(L+M)/2}\left|\mathbf{S}\right|^{-1/2}\left|\mathbf{P}\right|^{-1/2}(2\pi)^{M/2}\left|\boldsymbol{\Omega}\right|^{1/2}\mathcal{N}(\mathbf{y};\omega,\boldsymbol{\Omega}) \\ &\quad\times\ (2\pi)^{L/2}\left|\boldsymbol{\Lambda}\right|^{1/2}\mathcal{N}(\mathbf{x};\lambda,\boldsymbol{\Lambda})\,. \qquad (204)\end{aligned}$$

Multiplying out the 2π terms and collecting the determinant terms gives

$$\eta(\mathbf{y},\mathbf{x}) = c\,\mathcal{N}(\mathbf{y};\omega,\boldsymbol{\Omega})\;\mathcal{N}(\mathbf{x};\lambda,\boldsymbol{\Lambda})\,, \qquad (205)$$

where

$$c = \left\{\left|\mathbf{S}\right|^{-1}\cdot\left|\mathbf{P}\right|^{-1}\cdot\left|\boldsymbol{\Omega}\right|\cdot\left|\boldsymbol{\Lambda}\right|\right\}^{1/2}. \qquad (206)$$

Equation (205) is the desired expression if it can be shown that $c = 1$. This equality is demonstrated by using a theorem for block matrices [46, 49], which states that the determinant of the matrix

$$\mathbf{B} = \begin{bmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix} \qquad (207)$$

with nonsingular $\mathbf{B}_{11}$ and $\mathbf{B}_{22}$ is given by

$$\left|\mathbf{B}\right| = \left|\mathbf{B}_{11}\right|\cdot\left|\mathbf{B}_{22}-\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\right| \qquad (208)$$

$$= \left|\mathbf{B}_{22}\right|\cdot\left|\mathbf{B}_{11}-\mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21}\right|. \qquad (209)$$

First, the definitions of $\boldsymbol{\Omega}$ and $\boldsymbol{\Lambda}$ and equation (209) are noted to obtain

$$\left|\boldsymbol{\Omega}\right|\cdot\left|\boldsymbol{\Lambda}\right| = \left|\boldsymbol{\Sigma}\right|\cdot\left|\mathbf{P}-\mathbf{P}\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{P}\right| = \det\begin{bmatrix}\mathbf{P} & \mathbf{P}\mathbf{F}^{\mathrm{T}}\\ \mathbf{F}\mathbf{P} & \boldsymbol{\Sigma}\end{bmatrix}. \qquad (210)$$

It is also noted that

$$\left|\mathbf{P}^{-1}\right|\cdot\left|\mathbf{S}^{-1}\right| = \det\begin{bmatrix}\mathbf{P}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{S}^{-1}\end{bmatrix}. \qquad (211)$$

Given these two partitioned matrices, the desired product is

$$\left|\mathbf{S}\right|^{-1} \cdot \left|\mathbf{P}\right|^{-1} \cdot \left|\Omega\right| \cdot \left|\Lambda\right| = \left|\Gamma\right|, \tag{212}$$

where

$$\Gamma = \begin{bmatrix} \mathbf{P}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{P} & \mathbf{P}\mathbf{F}^{\mathrm{T}} \\ \mathbf{F}\mathbf{P} & \Sigma \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{F}^{\mathrm{T}} \\ \mathbf{S}^{-1}\mathbf{F}\mathbf{P} & \mathbf{S}^{-1}\Sigma \end{bmatrix}. \tag{213}$$

Application of equation (208) then yields

$$\begin{aligned} \left|\Gamma\right| &= \left|\mathbf{I}\right| \cdot \left|\mathbf{S}^{-1}\Sigma - \mathbf{S}^{-1}\mathbf{F}\mathbf{P}\mathbf{I}\mathbf{F}^{\mathrm{T}}\right| \\ &= \left|\mathbf{S}^{-1}\left(\Sigma - \mathbf{F}\mathbf{P}\mathbf{F}^{\mathrm{T}}\right)\right| \\ &= \left|\mathbf{S}^{-1}\mathbf{S}\right| = 1, \end{aligned} \tag{214}$$

which completes the proof.

APPENDIX B
HGMM JOINT STATE DENSITY

This appendix derives the conditional joint density of time-adjacent HGMM states, as given in equations (114) through (116). The starting point for this derivation is equation (111), which states that

$$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) \;=\; \mathcal{N}\left(\mathbf{x}_n; \mu_{n|N}, \mathbf{P}_{n|N}\right) \mathcal{N}(\mathbf{x}_{n-1}; \lambda_n, \mathbf{\Lambda}_n)\,, \tag{215}$$

where

$$\begin{aligned} \lambda_n &= (\mathbf{I} - \mathbf{H}_n\mathbf{A}_n)\,\mu_{n-1|n-1} + \mathbf{H}_n\mathbf{x}_n \\ \mathbf{\Lambda}_n &= (\mathbf{I} - \mathbf{H}_n\mathbf{A}_n)\,\mathbf{P}_{n-1|n-1} \\ \mathbf{H}_n &= \mathbf{P}_{n-1|n-1}\mathbf{A}_n^{\mathrm{T}}\mathbf{P}_{n|n-1}^{-1}. \end{aligned}$$

Substituting the functional forms of the normal densities gives

$$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) \;=\; (2\pi)^{-L}\left|\mathbf{P}_{n|N}\right|^{-1/2}\left|\mathbf{\Lambda}_n\right|^{-1/2}\exp\left\{-\frac{1}{2}\,\xi(\mathbf{x}_n, \mathbf{x}_{n-1})\right\}, \tag{216}$$

where

$$\xi(\mathbf{x}_n, \mathbf{x}_{n-1}) \;=\; \left(\mathbf{x}_n - \mu_{n|N}\right)^{\mathrm{T}}\mathbf{P}_{n|N}^{-1}\left(\mathbf{x}_n - \mu_{n|N}\right) + (\mathbf{x}_{n-1} - \lambda_n)^{\mathrm{T}}\,\mathbf{\Lambda}_n^{-1}\,(\mathbf{x}_{n-1} - \lambda_n). \tag{217}$$

Defining the conditional state error vector $\varepsilon_n = \mathbf{x}_n - \mu_{n|N}$ and recalling that

$$\begin{aligned} \mu_{n|n-1} &= \mathbf{A}_n\,\mu_{n-1|n-1} \\ \mu_{n-1|N} &= \mu_{n-1|n-1} + \mathbf{H}_n\left(\mu_{n|N} - \mu_{n|n-1}\right) \end{aligned}$$

allows λ_n to be rewritten as

$$\begin{aligned} \lambda_n &= \mu_{n-1|n-1} + \mathbf{H}_n\left(\mathbf{x}_n - \mathbf{A}_n\,\mu_{n-1|n-1}\right) \\ &= \mu_{n-1|n-1} + \mathbf{H}_n\left(\mu_{n|N} - \mu_{n|n-1}\right) + \mathbf{H}_n\,\varepsilon_n \\ &= \mu_{n-1|N} + \mathbf{H}_n\,\varepsilon_n\,. \end{aligned} \tag{218}$$

Defining the error vector

$$\mathbf{e}_{n-1} \;=\; \mathbf{x}_{n-1} - \lambda_n \;=\; \varepsilon_{n-1} - \mathbf{H}_n\,\varepsilon_n \tag{219}$$

then results in $\xi(\mathbf{x}_n, \mathbf{x}_{n-1})$ being expressed as

$$\begin{aligned} \xi(\mathbf{x}_n, \mathbf{x}_{n-1}) &= \varepsilon_n{}^{\mathrm{T}}\,\mathbf{P}_{n|N}^{-1}\,\varepsilon_n + \mathbf{e}_{n-1}{}^{\mathrm{T}}\,\mathbf{\Lambda}^{-1}\,\mathbf{e}_{n-1} \\ &= \varepsilon_n{}^{\mathrm{T}}\,\mathbf{P}_{n|N}^{-1}\,\varepsilon_n + (\varepsilon_{n-1} - \mathbf{H}_n\,\varepsilon_n)^{\mathrm{T}}\,\mathbf{\Lambda}^{-1}\,(\varepsilon_{n-1} - \mathbf{H}_n\,\varepsilon_n) \\ &= \varepsilon_n{}^{\mathrm{T}}\left(\mathbf{P}_{n|N}^{-1} + \mathbf{H}_n^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{H}_n\right)\varepsilon_n - \varepsilon_{n-1}^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{H}_n\,\varepsilon_n \\ &\qquad - \varepsilon_n^{\mathrm{T}}\mathbf{H}_n^{\mathrm{T}}\mathbf{\Lambda}_n^{-1}\varepsilon_{n-1} + \varepsilon_{n-1}{}^{\mathrm{T}}\,\mathbf{\Lambda}^{-1}\,\varepsilon_{n-1}. \end{aligned} \tag{220}$$

Equation (220) is written in matrix notation as

$$\xi(\mathbf{x}_n, \mathbf{x}_{n-1}) = \begin{bmatrix} \varepsilon_n \\ \varepsilon_{n-1} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{P}_{n|N}^{-1} + \mathbf{H}_n^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \mathbf{H}_n & -\mathbf{H}_n^{\mathrm{T}} \boldsymbol{\Lambda}_n^{-1} \\ -\boldsymbol{\Lambda}_n^{-1} \mathbf{H}_n & \boldsymbol{\Lambda}_n^{-1} \end{bmatrix} \begin{bmatrix} \varepsilon_n \\ \varepsilon_{n-1} \end{bmatrix}. \quad (221)$$

The joint density can then be written as

$$\gamma(\mathbf{x}_n, \mathbf{x}_{n-1}) = \mathcal{N}\left(\mathbf{x}_{[n,n-1]}; \mu_{[n,n-1]|N}, \mathbf{P}_{[n,n-1]|N}\right), \quad (222)$$

where the joint state variable is

$$\mathbf{x}_{[n,n-1]} = \begin{bmatrix} \mathbf{x}_n. \\ \mathbf{x}_{n-1} \end{bmatrix} \quad (223)$$

and the mean and covariance are

$$\mu_{[n,n-1]|N} = \begin{bmatrix} \mu_{n|N} \\ \mu_{n-1|N} \end{bmatrix} \quad (224)$$

$$\mathbf{P}_{[n,n-1]|N} = \begin{bmatrix} \mathbf{P}_{n|N}^{-1} + \mathbf{H}_n^{\mathrm{T}} \boldsymbol{\Lambda}_n^{-1} \mathbf{H}_n & -\mathbf{H}_n^{\mathrm{T}} \boldsymbol{\Lambda}_n^{-1} \\ -\boldsymbol{\Lambda}_n^{-1} \mathbf{H}_n & \boldsymbol{\Lambda}_n^{-1} \end{bmatrix}^{-1}, \quad (225)$$

respectively. There are no scale factors in equation (222) because, using an argument similar to that given in appendix A, it can be shown that

$$\left|\mathbf{P}_{n|N}\right| \cdot \left|\boldsymbol{\Lambda}_n\right| \cdot \left|\mathbf{P}_{[n,n-1]|N}^{-1}\right| = 1. \quad (226)$$

A more useful expression for the joint covariance matrix is obtained by applying the block matrix inversion theorem [49]. This theorem states that the inverse of block matrix $\mathbf{S}$ with nonsingular diagonal blocks $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ is

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} = \mathbf{S}^{-1} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}^{-1}, \quad (227)$$

where

$$\mathbf{T}_{11} = \mathbf{S}_{11\cdot 2}^{-1} \quad (228)$$

$$\mathbf{T}_{12} = -\mathbf{S}_{11}^{-1} \mathbf{S}_{12} \mathbf{S}_{22\cdot 1}^{-1} \quad (229)$$

$$\mathbf{T}_{21} = -\mathbf{S}_{22}^{-1} \mathbf{S}_{21} \mathbf{S}_{11\cdot 2}^{-1} \quad (230)$$

$$\mathbf{T}_{22} = \mathbf{S}_{22\cdot 1}^{-1}. \quad (231)$$

Here, $\mathbf{S}_{11\cdot 2}$ and $\mathbf{S}_{22\cdot 1}$ are the Schur complements

$$\begin{aligned} \mathbf{S}_{11\cdot 2} &= \mathbf{S}_{11} - \mathbf{S}_{21} \mathbf{S}_{22}^{-1} \mathbf{S}_{12} \\ \mathbf{S}_{22\cdot 1} &= \mathbf{S}_{22} - \mathbf{S}_{12} \mathbf{S}_{11}^{-1} \mathbf{S}_{21}. \end{aligned} \quad (232)$$

The structure of the joint covariance matrix dictates that the diagonal blocks should equal the covariance matrices $\mathbf{P}_{n|N}$ and $\mathbf{P}_{n-1|N}$. This outcome is confirmed by substituting $\mathbf{S}_{11} = \mathbf{P}_{n|N}^{-1} + \mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n^{-1}\mathbf{H}_n$, $\mathbf{S}_{12} = -\mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n^{-1}$, $\mathbf{S}_{21} = \boldsymbol{\Lambda}_n^{-1}\mathbf{H}_n$, and $\mathbf{S}_{22} = \boldsymbol{\Lambda}_n^{-1}$, so that the Schur complements are

$$\mathbf{S}_{11\cdot 2} = \left(\mathbf{P}_{n|N}^{-1} + \mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n^{-1}\mathbf{H}_n - \mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\mathbf{H}_n\right)^{-1} = \mathbf{P}_{n|N}^{-1} \tag{233}$$

$$\mathbf{S}_{22\cdot 1} = \boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}\left(\mathbf{P}_{n|N} + \mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbf{H}_n\right)^{-1}\mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}^{-1} = \left(\boldsymbol{\Lambda} + \mathbf{H}_n\mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}\right)^{-1}. \tag{234}$$

With these results, the first diagonal block of the joint covariance matrix is $\mathbf{T}_{11} = \mathbf{P}_{n|N}$, as it should be. The second diagonal block is

$$\begin{aligned}
\mathbf{T}_{22} &= \boldsymbol{\Lambda} + \mathbf{H}_n\mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}} \\
&= (\mathbf{I} - \mathbf{H}_n\mathbf{A}_n)\,\mathbf{P}_{n-1|n-1} + \mathbf{H}_n\mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}} \\
&= \mathbf{P}_{n-1|n-1} - \mathbf{H}_n\mathbf{A}_n\mathbf{P}_{n-1|n-1} + \mathbf{H}_n\mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}} \\
&= \mathbf{P}_{n-1|n-1} - \mathbf{H}_n\mathbf{P}_{n|n-1}\mathbf{P}_{n|n-1}^{-1}\mathbf{A}_n\mathbf{P}_{n-1|n-1} + \mathbf{H}_n\mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}} \\
&= \mathbf{P}_{n-1|n-1} - \mathbf{H}_n\mathbf{P}_{n|n-1}\mathbf{H}_n^{\mathrm{T}} + \mathbf{H}_n\mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}} \\
&= \mathbf{P}_{n-1|n-1} - \mathbf{H}_n\left(\mathbf{P}_{n|N} - \mathbf{P}_{n|n-1}\right)\mathbf{H}_n^{\mathrm{T}},
\end{aligned} \tag{235}$$

which is just the definition of $\mathbf{P}_{n-1|N}$ from the RTS smoother. The lower off-diagonal is given by

$$\mathbf{T}_{21} = \boldsymbol{\Lambda}_n\left(\boldsymbol{\Lambda}_n^{-1}\mathbf{H}_n\right)\mathbf{P}_{n|N} = \mathbf{H}_n\mathbf{P}_{n|N}. \tag{236}$$

Since the inverse of a symmetric matrix must also be symmetric, the cross-covariance matrix is known to be

$$\mathbf{P}_{n,n-1|N} = \mathbf{T}_{12} = \mathbf{T}_{21}^{\mathrm{T}} = \mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}. \tag{237}$$

To verify the validity of this expression, $\mathbf{T}_{12}$ is evaluated from the terms in the block matrix inverse, giving the expected result as

$$\begin{aligned}
\mathbf{T}_{12} &= \left(\mathbf{P}_{n|N}^{-1} + \mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n^{-1}\mathbf{H}_n\right)^{-1}\mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n^{-1}\mathbf{P}_{n-1|N} \\
&= \left[\mathbf{P}_{n|N} - \mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}\left(\boldsymbol{\Lambda} + \mathbf{H}_n\mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}\right)^{-1}\mathbf{H}_n\mathbf{P}_{n|N}\right]\mathbf{H}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n^{-1}\mathbf{P}_{n-1|N} \\
&= \mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_{n-1|N}^{-1}\mathbf{H}_n\mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}\right)\boldsymbol{\Lambda}_n^{-1}\mathbf{P}_{n-1|N} \\
&= \mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}\left[\mathbf{I} - \mathbf{P}_{n-1|N}^{-1}\left(\mathbf{P}_{n-1|N} - \boldsymbol{\Lambda}_n\right)\right]\boldsymbol{\Lambda}_n^{-1}\mathbf{P}_{n-1|N} \\
&= \mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}\left(\mathbf{I} - \mathbf{I} + \mathbf{P}_{n-1|N}^{-1}\boldsymbol{\Lambda}_n\right)\boldsymbol{\Lambda}_n^{-1}\mathbf{P}_{n-1|N} \\
&= \mathbf{P}_{n|N}\mathbf{H}_n^{\mathrm{T}}.
\end{aligned}$$

APPENDIX C
HGMM PARAMETER INVARIANCE PROPERTY

The objective is to show that the model with time-invariant parameters

$$\Theta = \{\mathbf{A}, \mathbf{B}, \mathbf{Q}, \mathbf{R}, \mu_0, \mathbf{P}_0\} \tag{238}$$

is equivalent under the measurement likelihood function to parameter set

$$\tilde{\Theta} = \left\{\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{Q}}, \mathbf{R}, \tilde{\mu}_0, \tilde{\mathbf{P}}_0\right\}, \tag{239}$$

where the transformed parameters satisfy

$$\tilde{\mathbf{A}} = \mathbf{U}_A \mathbf{A} \mathbf{U}_A^{-1} \tag{240}$$
$$\tilde{\mathbf{B}} = \mathbf{B} \mathbf{U}_0^{-1} \mathbf{U}_A^{-1} \tag{241}$$
$$\tilde{\mathbf{Q}} = \mathbf{U}_A \mathbf{U}_0 \mathbf{Q} \mathbf{U}_0^{\mathrm{T}} \mathbf{U}_A^{\mathrm{T}} \tag{242}$$
$$\tilde{\mu}_0 = \mathbf{U}_A \mathbf{U}_0 \mu_0 \tag{243}$$
$$\tilde{\mathbf{P}}_0 = \mathbf{U}_A \mathbf{U}_0 \mathbf{P}_0 \mathbf{U}_0^{\mathrm{T}} \mathbf{U}_A^{\mathrm{T}}. \tag{244}$$

Here, $\mathbf{U}_A$ can be any nonsingular $L \times L$ matrix and $\mathbf{U}_0$ can be any nonsingular $L \times L$ matrix that commutes with $\mathbf{A}$.

The invariance of the likelihood is demonstrated by showing the invariance of the measurement innovation ν_n and measurement-error covariance $\mathbf{\Sigma}_n$. Note that the state variables are *not* invariant to the parameter transformations and that they undergo a corresponding set of transformations defined by

$$\tilde{\mu}_{n|n-1} = \mathbf{U}_A \mathbf{U}_0 \mu_{n|n-1}$$
$$\tilde{\mu}_{n|n} = \mathbf{U}_A \mathbf{U}_0 \mu_{n|n}$$
$$\tilde{\mathbf{P}}_{n|n-1} = \mathbf{U}_A \mathbf{U}_0 \mathbf{P}_{n|n-1} \mathbf{U}_0^{\mathrm{T}} \mathbf{U}_A^{\mathrm{T}}$$
$$\tilde{\mathbf{P}}_{n|n} = \mathbf{U}_A \mathbf{U}_0 \mathbf{P}_{n|n} \mathbf{U}_0^{\mathrm{T}} \mathbf{U}_A^{\mathrm{T}}.$$

First, it is shown that output calculations with the transformed parameters and state variables produce the same measurement innovation and error covariance as the original model. It is then shown that the Baum recursions actually propagate these variables.

The measurement innovation is given in terms of the transformed states and model parameters as

$$\begin{aligned}\nu_n &= \mathbf{z}_n - \tilde{\mathbf{B}}\,\tilde{\mu}_{n|n-1} \\ &= \mathbf{z}_n - \mathbf{B}\,\mathbf{U}_0^{-1}\,\mathbf{U}_A^{-1}\mathbf{U}_A\,\mathbf{U}_0\,\mu_{n|n-1} \\ &= \mathbf{z}_n - \mathbf{B}\,\mu_{n|n-1} \\ &= \nu_n\,, \end{aligned} \tag{245}$$

and the measurement-error covariance matrix is

$$\begin{aligned}\tilde{\Sigma}_n &= \mathbf{R} + \tilde{\mathbf{B}}\tilde{\mathbf{P}}_{n|n-1}\tilde{\mathbf{B}}^{\mathrm{T}} \\ &= \mathbf{R} + \mathbf{U}_0^{-1}\,\mathbf{U}_A^{-1}\mathbf{U}_A\,\mathbf{U}_0\,\mathbf{P}_{n|n-1}\mathbf{U}_0^{\mathrm{T}}\,\mathbf{U}_A^{\mathrm{T}}\mathbf{U}_A^{-\mathrm{T}}\,\mathbf{U}_0^{-\mathrm{T}}\,\mathbf{B}^{\mathrm{T}} \\ &= \mathbf{R} + \mathbf{B}\mathbf{P}_{n|n-1}\mathbf{B}^{\mathrm{T}} \\ &= \Sigma_n. \end{aligned} \tag{246}$$

Thus, if it is assumed that the transformed model propagates the transformed state variables in the forward probability recursions, the invariance is demonstrated. The propagation of these variables is demonstrated by induction. The transformed variables fit the initial state conditions by definition, so that only the recursions need be examined. For the time update, it is observed that

$$\begin{aligned}\tilde{\mu}_{n|n-1} &= \tilde{\mathbf{A}}_n\,\tilde{\mu}_{n-1|n-1} \\ &= \mathbf{U}_A\,\mathbf{A}\,\mathbf{U}_A^{-1}\,\mathbf{U}_A\,\mathbf{U}_0\,\mu_{n-1|n-1} \\ &= \mathbf{U}_A\,\mathbf{A}\,\mathbf{U}_0\,\mu_{n-1|n-1} \\ &= \mathbf{U}_A\,\mathbf{U}_0\,\mathbf{A}\,\mu_{n-1|n-1} \\ &= \mathbf{U}_A\,\mathbf{U}_0\,\mu_{n|n-1}. \end{aligned}$$

The recursion for the transformed covariance matrix is

$$\begin{aligned}\tilde{\mathbf{P}}_{n|n-1} &= \tilde{\mathbf{Q}} + \tilde{\mathbf{A}}\tilde{\mathbf{P}}_{n-1|n-1}\tilde{\mathbf{A}}^{\mathrm{T}} \\ &= \mathbf{U}_A\,\mathbf{U}_0\,\mathbf{Q}\,\mathbf{U}_0^{\mathrm{T}}\,\mathbf{U}_A^{\mathrm{T}} + \mathbf{U}_A\,\mathbf{A}\,\mathbf{U}_A^{-1}\mathbf{U}_A\,\mathbf{U}_0\,\mathbf{P}_0\,\mathbf{U}_0^{\mathrm{T}}\,\mathbf{U}_A^{\mathrm{T}}\left(\mathbf{U}_A\,\mathbf{A}\,\mathbf{U}_A^{-1}\right)^{\mathrm{T}} \\ &= \mathbf{U}_A\,\mathbf{U}_0\left(\mathbf{Q} + \mathbf{A}\,\mathbf{P}_{n-1|n-1}\mathbf{A}^{\mathrm{T}}\right)\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_A^{\mathrm{T}} \\ &= \mathbf{U}_A\,\mathbf{U}_0\,\mathbf{P}_{n-1|n-1}\,\mathbf{U}_0^{\mathrm{T}}\,\mathbf{U}_A^{\mathrm{T}}. \end{aligned} \tag{247}$$

These expressions confirm that the time updates propagate the transformed variables.

Verification of the correct variable propagation during the measurement update begins by examining the gain matrix, which is given as

$$\begin{aligned}
\tilde{\mathbf{G}}_n &= \tilde{\mathbf{P}}_{n|n-1}\tilde{\mathbf{B}}^{\mathrm{T}}\tilde{\mathbf{\Sigma}}_n \\
&= \mathbf{U}_A\mathbf{U}_0\mathbf{P}_{n|n-1}\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_A^{\mathrm{T}}\left(\mathbf{B}\mathbf{U}_0^{-1}\mathbf{U}_A^{-1}\right)^{\mathrm{T}}\mathbf{\Sigma}_n \\
&= \mathbf{U}_A\mathbf{U}_0\mathbf{P}_{n|n-1}\mathbf{B}\mathbf{\Sigma}_n \\
&= \mathbf{U}_A\mathbf{U}_0\mathbf{G}_n. \qquad (248)
\end{aligned}$$

The measurement update for the mean is then

$$\begin{aligned}
\tilde{\mu}_{n|n} &= \left(\mathbf{I}-\tilde{\mathbf{G}}_n\tilde{\mathbf{B}}\right)\tilde{\mu}_{n|n-1}+\tilde{\mathbf{G}}_n\mathbf{z}_n \\
&= \left(\mathbf{I}-\mathbf{U}_A\mathbf{U}_0\mathbf{G}_n\mathbf{B}\mathbf{U}_0^{-1}\mathbf{U}_A^{-1}\right)\mathbf{U}_A\mathbf{U}_0\mu_{n|n-1}+\mathbf{U}_A\mathbf{U}_0\mathbf{G}_n\mathbf{z}_n \\
&= \mathbf{U}_A\mathbf{U}_0\left[\left(\mathbf{I}-\mathbf{G}_n\mathbf{B}\right)\mu_{n|n-1}+\mathbf{G}_n\mathbf{z}_n\right] \\
&= \mathbf{U}_A\mathbf{U}_0\mu_{n|n}, \qquad (249)
\end{aligned}$$

and the update for the covariance is

$$\begin{aligned}
\tilde{\mathbf{P}}_{n|n} &= \left(\mathbf{I}-\tilde{\mathbf{G}}_n\tilde{\mathbf{B}}\right)\tilde{\mathbf{P}}_{n|n-1} \\
&= \left(\mathbf{I}-\mathbf{U}_A\mathbf{U}_0\mathbf{G}_n\mathbf{B}\mathbf{U}_0^{-1}\mathbf{U}_A^{-1}\right) \\
&= \mathbf{U}_A\mathbf{U}_0\left[\left(\mathbf{I}-\tilde{\mathbf{G}}_n\tilde{\mathbf{B}}\right)\tilde{\mathbf{P}}_{n|n-1}\right]\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_A^{\mathrm{T}} \\
&= \mathbf{U}_A\mathbf{U}_0\mathbf{P}_{n|n}\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_A^{\mathrm{T}}. \qquad (250)
\end{aligned}$$

These expressions demonstrate that the measurement updates propagate the transformed variables, thereby completing the invariance proof.

REFERENCES

1. D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, Inc., New York, 1992.
2. P. M. Baggenstoss, "Class-Specific Feature Sets in Classification," *IEEE Transactions on Signal Processing*, vol. 47, no. 12, December 1999, pp. 3428–2432.
3. Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, Inc., New York, 1988.
4. A. E. Bryson and Y.-C. Ho, *Applied Optimal Control*, revised printing, Hemisphere Publishing Corporation, New York, 1975.
5. B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1979.
6. T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, Inc., Upper Saddle River, NJ, 2000.
7. L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1993.
8. F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, 1997.
9. L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Mathematical Statistics*, vol. 41, no. 1, 1970, pp. 164–171.
10. L. E. Baum and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Annals of Mathematical Statistics*, vol. 37, 1966, pp. 1554–1563.
11. L. E. Baum and J. A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology," *Bulletin of the American Meteorological Society*, vol. 73, 1967, pp. 360–363.
12. L. E. Baum, "An Inequality and Associate Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process," *Inequalities*, vol. 3, 1972, pp. 1–8.
13. A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotic Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, 1967, pp. 260–269.

14. G. D. Forney, Jr., "The Viterbi Algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, March 1973, pp. 268–278.

15. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood Estimation from Incomplete Data," *Journal of the Royal Statistical Society (B)*, vol. 39, no. 1, 1977, pp. 1–38.

16. W. J. Heiser, "Convergent Computation by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis," in *Recent Advances in Descriptive Multivariate Analysis*, W. J. Krzanowski, ed., Clarendon Press, Oxford, 1995.

17. J. D. Ferguson, "Hidden Markov Analysis: An Introduction," in *Proceedings of the IDA-CDR Symposium on Applications of Hidden Markov Models to Text and Speech*, J. D. Ferguson, ed., Princeton, NJ, October 1980.

18. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell System Technical Journal*, vol. 62, no. 4, April 1983, pp. 1035–1074.

19. L. A. Liporace, "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," *IEEE Transactions on Information Theory*, vol. 28, no. 5, September 1982, pp. 729–734.

20. B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains," *IEEE Transactions on Information Theory*, vol. 32, no. 2, March 1986, pp. 307–309.

21. A. B. Poritz, "Linear Predictive Hidden Markov Models and the Speech Signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1982, pp. 1291–1294.

22. B.-H. Juang and L. R. Rabiner, "Mixture Autoregressive Hidden Markov Models for Speech Signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 6, December 1985, pp. 1404–1413.

23. P. Kenny, M. Lennig, and P. Mermelstein, "A Linear Predictive HMM for Vector-Valued Observations with Applications to Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, February 1990, pp. 220–225.

24. L. Deng, M. Aksmanovic, X. Sun, and C. F. J. Wu, "Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary

States," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, October 1994, pp. 507–520.

25. V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML Estimation of a Stochastic Linear System with the EM Algorithm and Its Application to Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, October 1993, pp. 431–442.
26. J. D. Ferguson, "Variable Duration Models for Speech," in *Proceedings of the IDA-CDR Symposium on Applications of Hidden Markov Models to Text and Speech*, J. D. Ferguson, ed., Princeton, NJ, October 1980.
27. M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, September 1996, pp. 360–378.
28. T. Kailath, "A View of Three Decades of Linear Filtering Theory," *IEEE Transactions on Information Theory*, vol. 20, no. 2, October 1974, pp. 145–181.
29. R. L. Kashyap, "Maximum Likelihood Identification of Stochastic Linear Systems," *IEEE Transactions on Automatic Control*, vol. 15, no. 1, February 1970, pp. 25–34.
30. N. K. Gupta and R. K. Mehra, "Computational Aspects of Maximum Likelihood Estimation and Reduction in Sensitivity Function Calculations," *IEEE Transactions on Automatic Control*, vol. 19, December 1974, pp. 774–783.
31. R. H. Shumway and D. S. Stoffer, "An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, 1982, pp. 253–264.
32. M. W. Watson and R. F. Engle, "Alternative Algorithms for the Estimation of Dynamic Factor, MIMIC and Varying Coefficient Regression Models," *Journal of Econometrics*, vol. 23, 1983, pp. 385–400.
33. I. Ziskind and D. Hertz, "Maximum Likelihood Localization of Narrow-band Autoregressive Sources via the EM Algorithm," *IEEE Transactions on Signal Processing*, vol. 41, no. 8, August 1993, pp. 2719–2724.
34. E. Weinstein, A. V. Oppenheim, M. Feder, and J. R. Buck, "Iterative and Sequential Algorithms for Multisensor Signal Enhancement," *IEEE Transactions on Signal Processing*, vol. 42, no. 14, April 1994, pp. 846–859.
35. L. Deng and X. Shen, "Maximum Likelihood in Statistical Estimation of Dynamic Systems: Decomposition Algorithm and Simulation Results," *Signal Processing*, vol. 57, 1997, pp. 65–79.

36. R. J. Elliot and V. Krishnamurthy, "New Finite-Dimensional Filters for Parameter Estimation of Discrete-Time Linear Gaussian Models," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, May 1997, pp. 938–951.

37. H. W. Sorenson and D. L. Alspach, "Recursive Bayesian Estimation Using Gaussian Sums," *Automatica*, vol. 7, 1971, pp. 465–479.

38. J. T.-H. Lo, "Finite-Dimensional Sensor Orbits and Optimal Nonlinear Filtering," *IEEE Transactions on Information Theory*, vol. 18, no. 5, September 1972, pp. 583–588.

39. R. L. Streit, "The Relationship Between Kalman Filters and Infinite-State Hidden Markov Models," NUWC Technical Memorandum 921088, Naval Undersea Warfare Center Division, Newport, RI, 1992.

40. A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, Inc., New York, 1970.

41. H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum Likelihood Estimates of Linear Dynamic Systems," *AAIA Journal*, vol. 3, 1965, pp. 1445–1450.

42. D. Q. Mayne, "A Solution to the Smoothing Problem for Linear Dynamical Systems," *Automatica*, vol. 4, 1966, pp. 73–92.

43. D. C. Fraser and J. E. Potter, "The Optimum Linear Smoother as a Combination of Two Optimum Linear Filters," *IEEE Transactions on Automatic Control*, vol. 7, no. 4, August 1969, pp. 387–390.

44. F. C. Schweppe, "Evaluation of Likelihood Functions for Gaussian Signals," *IEEE Transactions on Information Theory*, vol. 11, 1965, pp. 61–70.

45. R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

46. F. A. Graybill, *Matrices with Applications in Statistics*, 2nd edition, Wadsworth, Inc., Pacific Grove, CA, 1983.

47. A. Graham, *Kronecker Products and Matrix Calculus with Applications*, John Wiley & Sons, Inc., New York, 1981.

48. Private communication with M. L. Graham, Naval Undersea Warfare Center Division, Newport, RI, January 2001.

49. L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time-Series Analysis*, Addison-Wesley Publishing Company, Inc., Reading, MA, 1991.

INITIAL DISTRIBUTION LIST

Addressee	No. of Copies
Office of Naval Research (ONR 321US — J. Tague)	1
Naval Sea Systems Command, ASTO/USW (SEA 93 — R. Zarnich)	1
Naval Research Laboratory (L. Couchman, T. C. Yang, T. Hayward)	3
Naval Postgraduate School (C. Therrien)	1
Defense Technical Information Center	2
Boston University (M. Ostendorf)	1
Duke University (L. Carin)	1
Johns Hopkins University (F. Jelinek)	1
Pennsylvania State University (D. Abraham)	1
Rice University (R. Baraniuk)	1
Stanford University (T. Kailath)	1
Technical University of Crete (V. Digalakis)	1
Tel-Aviv University (E. Weinstein)	1
Texas A&M University (N. Kehtarnavaz, E. Dougherty, D. Halverson, E. Parzen)	4
University of California, Los Angeles (A. Sayed)	1
University of Colorado, Boulder (L. Scharf)	1
University of Connecticut (P. Willet, Y. Bar-Shalom)	2
University of Massachusetts, Dartmouth (J. Buck)	1
University of Melbourne (V. Krishnamurthy)	1
University of Missouri, St. Louis (C. Chui)	1
University of Queensland (G. McLachlan)	1
University of Rhode Island (D. Tufts, S. Kay)	2
University of Washington (L. Atlas)	1
University of Waterloo (L. Deng)	1
Australian Academy of Science (B. D. O. Anderson)	1
AT&T Laboratories-Research (L. Rabiner)	1
Lockheed-Martin-Arizona (D. Kil)	1
Lucent Technologies-Bell Laboratories (B.-H. Juang, M. Sondhi)	2